Data Quality Assessment Approaches for Event-based Surveillance Systems

Mehtab Alam SYED†, Elena ARSEVSKA, Mathieu ROCHE, Maguelonne TEISSEIRE

UMR TETIS, CIRAD, Montpellier, France, UMR TETIS, INRAE, Montpellier, France

Keywords: Text Mining; Natural language processing (NLP); Data quality; Disease surveillance; EBS

Citation: SYED M.A., ARSEVSKA E., ROCHE M., et al.: Data Quality Assessment Approaches for Event-based Surveillance Systems. Data Intelligence, Vol. XX, pp. 1–28, Art. No.: 2025??XX, 2025. DOI: https://doi.org/10.3724/2096-7004.di.2025.0063

ABSTRACT

Online news sources are popular resources for learning about current health situations and developing event-based surveillance (EBS) systems. However, having access to diverse information originating from multiple sources can misinform stakeholders, eventually leading to false health risks. The existing literature contains several techniques for performing data quality evaluation to minimize the effects of misleading information. We mainly proposed three approaches to assess the quality of news sources. In our research, our primary focus was on ensuring data quality assessment at two levels: 1) News article level and 2) News source level. We explored data quality assessment at the news article level through two main approaches: 1) Data-driven score-based approach and 2) Metadata-based machine learning (ML) approach. The data-driven score-based approach aims to classify relevant and irrelevant news articles, adding an explainability aspect in the context of EBS. Similarly, the metadata approach is employed for classification, utilizing news article metadata features in ML models to highlight important metadata features. For source-level quality assessment, we identified exogenous metadata attributes such as source categorization and geographical coverage associated with news sources, extracting this information automatically. With the help of extracted source metadata, we conducted the classification of news sources. The obtained results hold significance in terms of prioritizing news sources within the context of EBS. Nevertheless, further investigation is required to enhance the methodology of this approach.

1. INTRODUCTION

Outbreaks of infectious diseases pose serious threats to public, animal, and plant health (one health) [1]. Moreover, infectious disease outbreaks affect not only one health but also the national and international

† Corresponding author: Mehtab Alam SYED (E-mail: mehtab_alam.syed@cirad.fr).

economy and trade [2]. Therefore, it is important to implement health surveillance methods to recognize potential infectious disease outbreaks and to minimize their associated devastating effects on affected population and indirectly on the society. In the existing literature [2-3], there are two main types of surveillance: 1) event-based surveillance (EBS) and 2) indicator-based surveillance (IBS). IBS uses official sources to detect important disease outbreaks [4]. They produce structured and reliable data, offering an extensive range of information regarding the pathogen, outbreak source, species, clinical signs, etc. As a result of official procedures, the declaration of outbreaks experiences a considerable time delay. Whereas, EBS refers to the collection of information regarding events that hold the potential to pose risks to public health reported in unstructured data sources like textual data, i.e., news articles, social media updates [5]. According to the World Health Organization (WHO), approximately 60% of all outbreaks are identified through informal sources [6]. These two surveillance strategies complement one another in terms of benefits due to their unique data collection, verification, assessment, and data interpretation processes [3] and are treated as the fundamental building blocks in constructing a comprehensive surveillance system [7]. Our research focused mainly on EBS, whereas IBS was beyond the scope of our study.

EBS is the organized process of detecting and reporting information on potential health threats and hazards (i.e. represented as events), most commonly as outbreak/cases, to healthcare authorities by rapid capturing of information from different unstructured data sources [7]. It enables health authorities to be better prepared for different disease outbreaks by functioning as a key component of an effective early warning system [3, 7]. For information acquisition, online information sources e.g. news articles, blogs, social media e.g., Twitter, and other ad-hoc reports e.g., access to laboratory reports, electronic health records, expert networks and exchanges are preferred in EBS systems [8-11] as compared to traditional data collection methods which are labor-intensive [12-13].

There are three types of EBS systems: moderated, partially moderated, and fully automated [14]. The way of flow of information in these EBS systems from online data sources, e.g. news aggregators, depicts its level of automation. The final output of all these types of EBS systems is to identify and extract signals or potential events (health threat from potential disease in a certain region over time) from heterogenous data sources [15]. In every type of EBS systems mentioned, there are certain advantages and disadvantages dependent on the level of priorities of certain factors. For instance, the Program for Monitoring Emerging Diseases (ProMED) is an example of a moderated system in which experts identify news articles, validate the content and report events [9, 16]. The main advantage of this system is less signal-to-noise ratio (low false outbreak detection rate) due to human validation of content, with disadvantages of resource limits (experts), situational awareness and expert biases towards the events. Similarly, the Global Public Health Intelligence Network (GPHIN) [10] is a partially moderated system that automatically identifies a stream of thousands of news articles per day and group of experts and further moderated by group of experts to identify events. It has the advantage of automated data collection method but with the same disadvantages as ProMED. Fully automated systems include the European Commission Medical Information System (MedISys) [17], Platform for Automated extraction of Disease Information (PADI-web) [18] from the web and HealthMap [8]. Unlike moderated systems, fully automated systems are faster at processing data and cost-efficient as compared to moderated systems. However, the main weakness of such systems

is the higher signal-to-noise ratio as compared to moderated systems, as well as less accuracy of event information i.e., there is significantly higher rate of identifying false health threats or information associated with health threats [19]. In our research, we focused on fully automated EBS system.

More than 60% of the first signals on new outbreaks come from news sources [20]. This finding underscores the role of news sources as early indicators of potential disease outbreaks, highlighting the need for data quality assurance in EBS systems. Online news information is diverse and collected from heterogeneous online data sources, it gets crucial to verify this unstructured information that can pose serious threats to public health to avoid misinformation (i.e. a piece of information that is false having no evidence) [21] and disinformation (i.e. intentionally generated false information) [22]. These sources are preferred as they improved timeliness and detection of outbreak-related information. To avoid false information on potential outbreaks, it is important to verify the information associated with the online news at two levels, i.e., news article, and news source in order to get accurate and reliable information.

The remainder of this article is structured as follows. In Section 2, we review the state-of-the-art approaches related to our work. Section 3 outlines the objectives of our research and highlights our key contributions. The datasets used for our experiments are described in Section 4. In Section 5, we explained data-driven score-based approach, followed by metadata-based machine learning classification method in Section 6. The results of our experiments are presented in Section 7. We provide a detailed discussion of these results in Section 8. Finally, Section 9 concludes the article and offers perspectives for future research.

2. STATE-OF-THE-ART

News articles of online news sources serve as vital data streams in disease surveillance systems, enabling real-time outbreak detection and informed public health interventions [23]. The aim is to leverage existing research in order to ensure the quality of data by evaluating the reliability, relevance, and accuracy of news articles and news sources, which serve as the primary information source for Early Warning Systems (EWS).

Research on data quality, which is crucial for evaluating online news sources and constructing EBS systems, began in the 1990s. Wang and Strong [24] defined data quality as "the information which is fit for use". The dimensions for assessing data quality are a set of attributes representing single or multiple aspects of data, including the currency, accuracy, relevance, authority, and purpose of information [25]. Data quality of news sources in the context of an EBS system refers to the reliability, relevance, timeliness, bias, and accuracy of the data collected and used for surveillance purposes [26]. Therefore, it is important to ensure these dimensions to avoid reducing the false alerts for disease surveillance. Reliability refers to the trustworthiness and credibility of the information being reported by news sources. Therefore, reputable news sources are bounded with strict fact-checking and verification processes before publishing information [27]. In the existing literature, there is a degree of overlap identified among the data quality dimensions and their assessment methods. For instance, Mandalios and Jane [28] used the following

assessment criteria to evaluate online sources: purpose, authority and credibility, accuracy and reliability, currency and timeliness and objectivity. In addition, Zhu and Gauch [29] proposed six quality metrics, including currency, availability, information-to-noise ratio, authority, popularity and cohesiveness, for investigating the assessment of online sources. Additionally, Nozato and Yoshiko [30] stated that the timeliness, depth, reputation, and accuracy of online sources are the most important data quality dimensions. Another study [31] used the quality attributes of the respondents and general perception of the news sources for news classification. Moreover, another study [32] investigated news credibility assessments by comparing crowds and expert opinions to understand the differentiation in the rating of the source. Relevant news sources offer information about better understanding of disease, mode of transmission, symptoms, level of risks in timely manner [33]. Accuracy of information from news sources depends on several factors, i.e., source credibility, transparency of information, cross-referencing across credible sources, bias, and experts input etc [34]. Moreover, these quality dimensions are dependent on multiple sub factors which are needed to be evaluated.

In addition to the data quality dimensions and their assessment methods as described above, there exist different studies that employ various state-of-the-art techniques [12] based on text mining, information retrieval, machine learning, deep learning and knowledge representation graphs for assessing the relevance of news sources. For example, Essam and Elsayed [35] defined a specialized information retrieval technique by assessing the topics and subtopics of the news to identify highly relevant background articles. Elhadad et al. [36] adopted a machine learning technique for extracting features from the news content and prepared a complex set of metadata for identifying the credibility of the news sources. Another study [37] proposed a method based on deep learning techniques to find patterns in news sources to avoid false information, rumors, spam, fake news, and disinformation. Moreover, Hu et al. [38] analysed the visual layout information of news homepages to utilize the mutual relationship that exists between news articles and news sources using a semi-supervised learning algorithm. However, this approach is not only based on a computationally expensive learning model to establish a relationship between new articles and sources but is also limited to small news corpora. To address this limitation, a system named Media Rank was designed [39] to incorporate large datasets for measuring the quality of news sources by a mix of computational signals reflecting peer reputation, reporting bias, bottom-line pressure, and popularity. A study employing the application of knowledge graphs by Rudnik et al. [40] implemented a method using a Wikidata knowledge base for generating the semantic annotation of news articles to filter relevant news articles. Additionally, Shu et al. [41] presented a fake news detection system using a combination of text and metadata analysis, highlighting the significance of user behavior patterns in verifying news credibility. Perez-Rosas et al. [42] developed a machine learning framework for automatically detecting fake news by focusing on linguistic features and psychological factors.

Metadata refers to structured information that provides details about various forms of data such as images, multimedia, books, and scientific articles [43-44]. In a research study, a metadata approach is used for categorization of historic newspaper collection. This metadata was collected by analysis of fined-grained search patterns within the newspaper collection [45]. In another research study, two primary types of metadata are introduced for digital news article archives [46]. These metadata categories consist

of explicit metadata (linked to news articles) and implicit metadata (crucial metadata embedded within the content), which serve the purpose of searching for news articles within archives. In another research, Bidirectional Encoder Representations from Transformers (BERT) model is proposed for the merging of text representations with metadata and knowledge graph embeddings, specifically encoding author-related information for book classification task [47]. In another research study, the DANIEL system, which is a text genre-based Information Extraction (IE) system, was proposed to efficiently filter out irrelevant documents in epidemic surveillance, particularly for low-resourced languages [48]. The benefit of this method was to increase coverage across a variety of languages at a low cost, rather than focusing solely on optimizing results for a specific language. In another research study, a framework is proposed for the profiling of cities through automatic extraction and analysis of metadata of news articles using data mining and machine learning techniques [49]. The cities profiles were characterized in terms of criminality, events, services, urban problems, decay, and accidents. Another research proposed a neural network based approach for multi-label document classification, in which two heterogeneous graphs are constructed i.e., metadata heterogeneous graph for modelling various types of metadata and their topological relations and label heterogeneous graph constructed based on labels hierarchy and their statistical dependencies [50]. In another research study, an approach was proposed for multi-label document classification using the available metadata for evaluating the performance of metadata-based features compared to content-based methods [51]. The proposed technique has been assessed for two diverse datasets, namely, from the Journal of universal computer science (JUCS) dataset and dataset of the articles published by the Association for computing machinery (ACM). Another research proposed transformer-based models for finding the documents that contain epidemic events and event extraction, with focus on high-resource and low-resource languages [52].

Recent studies have shown the importance of metadata in assessing the credibility and trustworthiness of news sources [53-54]. Ribeiro et al. (2018) explored media trust and use among urban news consumers in Brazil, analysing how ethical precepts connect to media credibility and trust [55]. Another study focused on the role of metadata in cognitive authority, examining environmental activists perceptions of media credibility [56]. Furthermore, Zhou and Zafarani (2020) investigated the utilization of user and content metadata for fake news detection, illustrating its potential in distinguishing between credible and non-credible sources [57]. Similarly, Baly et al. (2018) leveraged metadata features in their approach to fake news detection, emphasizing the role of source reliability [58]. These studies highlight the significance of metadata in determining the reliability of news sources and guiding.

A study [59] proposed that is based on a domain-oriented news article classification problem, is taken as fundamental to this research and used to develop the proposed approach. The research discussed a direct method (i.e., identification, review, and evaluation of known sources to find relevant information sources) and an indirect method that assess quality attributes of news content and metadata. To fill the research gap, we proposed the assessment of the data quality at two levels, i.e., 1) News article, 2) News Source. In the first step, we proposed automated method for extraction of quality attributes from the metadata and content of the news article and assessment of the data quality of news articles. The second step is to compliment the first step, we additionally identify

quality attributes associated with news source and the possible ones to extract automatically and evaluate the quality at News source level.

3. OBJECTIVES AND CONTRIBUTIONS

The main objective is to assess quality attributes to ensure the relevant information in EBS systems. The subsidiary research questions to support the main question are as follows:

1. What are the quality attributes of the news article identified from metadata and content, and how to evaluate these attributes for news article in EBS systems?
2. What are the external quality attributes to ensures the quality of news source, and evaluate these attributes to verify news source in EBS systems?
3. How the combination of attributes at two levels ensure the quality of information provided to EBS systems?

The following contributions included to address the objectives are as follows:

1. We proposed a data-driven score approach to assess the quality of online news articles in EBS. The online news articles are assessed through metadata and content of the news articles in order to filter relevant news articles.
2. We proposed machine learning approaches to classify relevant news articles in EBS through metadata features.
3. We proposed several quality attributes (source metadata) of news sources and their automated extraction. Moreover, we analysed the impact of news article metadata and source metadata towards classification of relevant news in EBS.

4. DATASETS

We annotated our data by expert having strong background in data science and epidemiology by keeping in mind the goal to ensure that data quality assessment was objective. In order to evaluate our contributions, we proposed two datasets. The details of these datasets are discussed in the subsequent sections.

4.1 EBS News Classification Dataset

This dataset consists of news articles related to Avian-Influenza (AI) events extract from PADI-web, with 317 articles classified as relevant and 374 articles classified as irrelevant. This dataset is manually labelled by epidemiologists for the news classification task. The dataset is limited in size because it involved manual annotation, which required human effort. Each entry in the dataset includes the following information: ID, title, text, URL, language, source language, creation date and class (label) of the news article. *Relevant* class contains AI outbreaks. While *irrelevant*, does not contain an event of AI outbreaks.

4.2 EBS Source Relevance Classification Dataset

The dataset contains the news sources detected by the PADI-web relevant (detected articles for avian-influenza outbreaks) or irrelevant with no event. This dataset is derived from the news article dataset in Section 4.1. Each entry in the dataset includes the following information: news_source, source_description, relevant_frequency, irrelevant_frequency, annotated_category, annotated_geographical_coverage and confidence (label) on the news source. For the experiments, confidence is a binary variable employed to assess the source classification task. The different approaches to assess the data quality in the context of EBS are discussed in the subsequent sections.

5. A DATA-DRIVEN SCORE-BASED APPROACH

Data quality measures (DQM) are the metrics to rank elements based on their quality, facilitating the identification of reliable news sources in terms of relevance, accuracy, and reputation [60]. Various criteria exist for computing the data quality of online news sources, including metadata attributes and attributes extracted from the content of the news article. Alomar et al. [59] proposed the measures of metadata score (MS) derived from extracted metadata, content score (CS) computed from extraction of various attributes inside the content of news article. Our approach proposed a new measure called epidemiological entity extraction score (E3S) calculated using weighted named-entities, specifically spatio-temporal entities related to epidemiology. We choose the unsupervised data-driven score approach for news classification because it helps us understand how different factors affect the results. Moreover, the dataset size and its focus on specific information can also be reasons for choosing this approach. The overall process pipeline, including all the components, is depicted in Figure 1.

The details of these measures are as follows:

1. **Metadata Score (MS):** Metadata plays a crucial role in rapidly retrieving information [61]. Search engines rely on metadata to quickly identify relevant results from countless online sources [61]. Reliable online sources define specific tags that are analysed and retrieved by search engines to deliver ranked results.
   When assessing the metadata of news sources, various metadata attributes are taken into account [59]. The different kinds of metadata of the news article are as follows:
   (a) **Title:** The title of the news article, which provides a brief summary or description of the story.
   (b) **Author:** The name of the individual or organization responsible for creating or reporting the news.
   (c) **Publication Date:** The date when the news article was published or made available to the public.
   (d) **Source:** The name of the news organization or publication that produced the article.
   (e) **Description:** A brief summary or abstract that provides an overview of the article's content.
   (f) **Keywords/Tags:** Relevant keywords or tags assigned to the article to describe its topic, subject, or themes. These help in categorization and searchability.
   (g) **Language:** The language in which the news article is written.

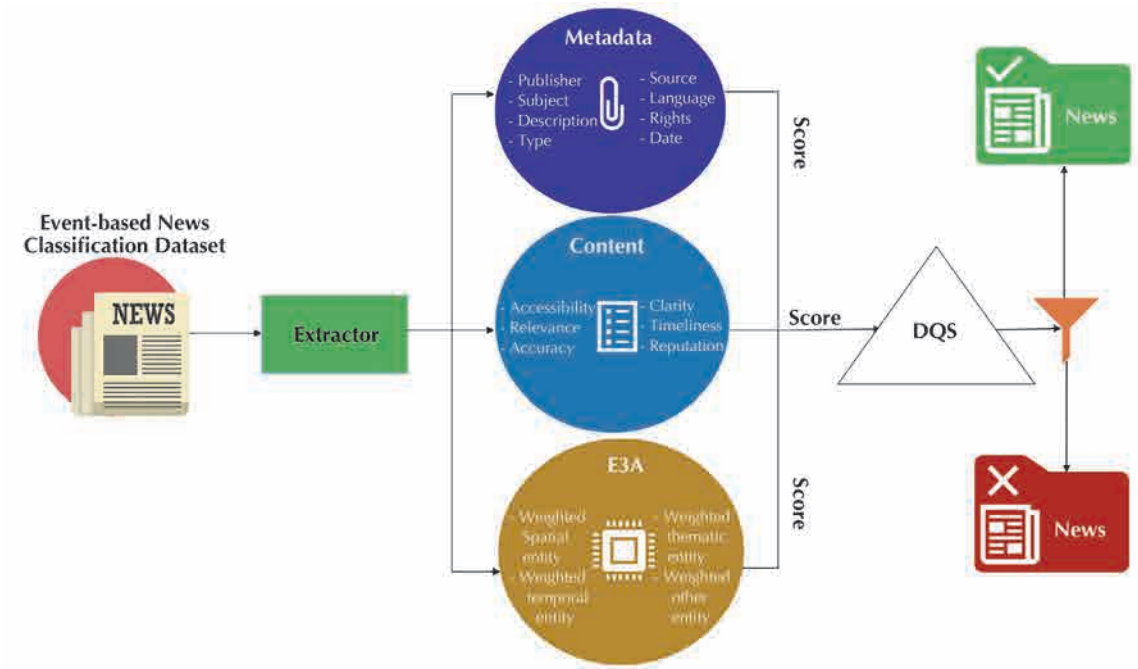

Figure 1. Process Pipeline: A data-driven score-based approach to assess data quality

(h) **type:** Type of the news article, e.g. topic or category.

(i) **Rights:** Copyright of the source.

We have identified and selected metadata attributes that are both relevant and widely accessible in most news articles, e.g., title (subject), description, author, date (publication date), type (topic of news article), rights (copyrights), source (URL of news article) and language. The extraction of each metadata attribute from the online news article allows for the computation of the metadata score (MS) using the formula proposed by [59]. MS is defined based on 8 metadata attributes, which is as follows:

$$MS = \sum_{n=1}^{8} Presence(attribute_n)$$
$$Presence(attribute_n) = \begin{cases} 1, & attribute \notin metadata \\ 2, & attribute \in metadata \end{cases} \quad (1)$$

We used the same scoring mechanism as discussed in the baseline approach. The scoring mechanism for the metadata attributes in this approach is, i.e., a score of 2 is assigned for the presence of an attribute, and a score of 1 is given for its absence.

2. **Content Score (CS):** Online news content comprises information about one or more events presented in the form of electronically available information for the public [62]. The quality of news articles is ensured through considerations of currency, timeliness, relevance, accuracy, and

impact [28]. Therefore, it is important to analyse the content of news sources and extract the quality attributes to quantify the data quality score.

The extraction of quality attributes from the content are achieved through an automated process. Various content attributes are taken into account for assessing the content of news sources [59]. The quality attributes selected for content assessment include accessibility, relevance, accuracy, clarity, timeliness, and reputation. **Accessibility** is the preliminary step of analysing content to access an online news source. Therefore, it is to ensure that the online news source is available and accessed without any barrier. Moreover, it is also possible that it is available but with the restricted access such that it is not possible to access with any browser or external tools. Despite, in some cases, it is also possible that the online news sources are unavailable for future use in digital form. **Relevance** in the baseline approach is proposed by identifying some epidemiological attributes i.e. affected hosts, agent that affects the host and the location of the affected host. This is not the same as the Relevance of the news article. We shortlist *spaCy* [63] natural language processing (NLP) python library is used to perform named entity recognition (NER) to extract locations, hosts and agents respectively. spaCy is easy to adapt and customize its models and components to suit specific NER requirements. This model can be utilized for improving the NER accuracy. Some examples of the hosts and agents of disease avian-influenza are (chicken, pigs, horse, duck, goose etc) and (H5N8, H5N1, highly pathogenic avian-influenza etc) respectively. **Accuracy** is dependent on the information provided by the news sources that are the facts that can be verified and validated. In the context of EBS, it could be that the news content provide information about any health risk, outbreak information, or it could be the number of cases respectively. Alternatively, poor relevance can have poor accuracy, but not vice versa. **Clarity** is the quality of being logical, consistent and completely understandable in terms of content that is similarly reflected in the metadata. Clarity of the article is poor if only the title is available in the metadata, and clarity is adequate if other metadata attributes are available [59]. A good clarity is if the subject, description, type etc. are available in the metadata of the news article. **Timeliness** is important to know if the content of the news article relates to the current context of the events. Otherwise, the claims may not be considered, or they may be wrongly interpreted. Timeliness is the time of an outbreak saved by detection in EBS relative to the onset of the outbreak [64]. Furthermore, Timeliness (days) is calculated by the following equation [65]:

$$Timeliness[days] = T_{alarm} - T_{onset} \tag{2}$$

where T_{alarm} is the time of the event reported in the event-based system and T_{onset} can be validated from the health information databases. Lastly, **Reputation** of news sources is extracted using *MediaRank* [39] algorithm which is calculated on multiple factors i.e. popularity, peer reputation, reporting bias and breadth and bottom-line pressure. E.g. the general reputation ranking using *MediaRank* [39] of *New York Times* is '1' and BBC is '5'. Therefore, the general reputation of the news source has the impact on the content quality, as it is computed on considering multiple factors. After the extraction of these attributes, the CS is computed using the following formulas in the baseline approach [59]:

$$CS = \sum_{n=1}^{6} Presence(attribute_n)$$
$$Presence(attribute_n) = \begin{cases} 1, \ Not\ available \\ 2, \ Partially\ available \\ 3, \ Available \end{cases} \quad (3)$$

We adapt the same scoring mechanism proposed by baseline approach. We shortlisted 6 attributes associated with the content ash shown in Figure 1. The interpretation of the CS is '1' means the attribute is not available, e.g., if the news article is not accessible. Moreover, '2' score represents the attribute is partially available, e.g., relevance is dependent on host, agent so if one of them is not available, then it is said to be partially available. Lastly, '3' score represents that attribute is completely available, e.g., if the news article is completely accessible online. Subsequently, the next step is the main contribution of this approach to extract the relevant contextual information from the online news article to enhance the baseline approach.

3. **Epidemiological Entity Extraction Score (E3S):** Event extraction and early warning detection are the key components of EBS [66]. An event is a verified set of processed epidemiological information of an outbreak [15]. It contains attributes such as location, occurrence date associated with epidemiological entities such as disease or unknown syndrome, symptoms, hosts, agents, etc. [15]. More precisely, this information is available in text in the form of spatio-temporal entities (when, where) and epidemiological entities (which) i.e. disease, host, agent, symptoms etc. Furthermore, these attributes are extracted from text using NLP techniques. The measure (E3S) is dependent on extracted spatial, temporal and epidemiological information within the news articles.
   In this measure, the title and content of a news article are processed and then named-entities are extracted. It is not sufficient to extract epidemiological (thematic) entities with state-of-the-art Name-Entity recognition (NER) techniques. In our approach, we categorized these entities into spatial, temporal, thematic and other entities. A rule-based approach is adapted to extend spaCy NER for extracting and classifying thematic entities such as hosts (e.g. humans, birds, pigs) that are associated with the disease and pathogens or agents e.g. H5N1, H5N8, HPAI, etc. spaCy was the first choice for NER task due to its balanced nature in terms of speed, accuracy, pretrained models, and ease of use as compared to other NLP libraries like NLTK, Stanford NLP. After extracting named entities from the title and content of news articles, weights are assigned to these categories depending on availability in title and content of the news article. It results into quantifying their epidemiological context in relation with its corresponding title and content. We named the resulting spatial, temporal and thematic entities as relevant entities in the context of a particular EBS (i.e. specific to proposed work) are termed as 'Contextual Entities (CE)'. For instance, *spatial entities* [67] are the names of the geographical or spatial location available in the text. Moreover, *temporal entities* [68] are the information of date, time and duration available in the text, *thematic entities* [69] are the information about health related terms in the text. whereas, the remaining identified entities are labelled as 'Non-contextual Entities (NE)'. The weights assigned to the entities are calculated by the following equations:

$$EntityWeight_n = \begin{cases} 2, & CE \in Title\ Sentence \\ 1.5, & CE \in Content \\ 1, & NE \end{cases} \qquad (4)$$

$$E3S = \sum_{i=1}^{n} CE\_Weight\ /\ E\_Weight$$

The weights are assigned based on two criteria i.e. 1) title and description of news articles, 2) types of entities. Double weights (i.e. 2) are assigned to these entities because of their occurrences in the title of the news articles, as title is the most important element of the news article, e.g., mostly event sentence is available in the title of the news article. Whereas, a weight of 1.5 is assigned to each CE entity based on their occurrences in the content of the news article. CE gives the contextual information of the event related information, so more weight is assigned as compared to other information in the news article. Lastly, a weight of 1 is assigned to each NE regardless of their occurrences in the title and content of the news articles. The E3S is calculated as the sum of CE_Weight to the sum of E_Weight (weight of all entities) in the news article, i.e. title and content. The resulting E3S provides weights of the news article in the epidemiological context having more chances to detect events.

The Word cloud visualization provides the most frequently used relevant words using different font sizes, indicating their occurrence frequency [70]. Figure 2 visualizes the occurrences of CEs in the Event-Based News Classification Dataset. We applied these score measures to classify the AI news dataset into relevant and irrelevant news. The results of these experiments are discussed in the subsequent section.

Figure 2. Word Cloud of Extracted CEs.

4. **Data Quality Score (DQS):** We used DQS for classifying the news article into relevant and irrelevant. The data quality score (DQS) is computed as the average of MS, CS and E3S.

6. METADATA-BASED MACHINE LEARNING CLASSIFICATION

To compliment the first approach, we proposed a machine learning approach to classify online news through metadata features. The two approaches differ in their methodology. In the data-driven approach, metadata features are used to score and filter relevant news articles. In contrast, the machine learning (ML) approach uses the same metadata features in ML models to identify which features are most relevant for filtering news articles in the EBS systems. This dual methodology allows for a comprehensive evaluation of metadata's role in both scoring and filtering processes. It is worthwhile to explore an additional perspective: the assessment and assurance of the external quality attributes associated with the sources of news data. This aspect is instrumental in understanding how news sources have historically contributed information to EBS systems and, in doing so, can significantly enhance the EBS system effectiveness. It serves as a mechanism to ensure that the information incorporated into these systems originates from reputable and dependable news sources.

In our contribution, we address mainly two types of metadata associated with the news: a) News article metadata and b) News source metadata. News article metadata is readily available and is directly associated with the news article, providing crucial information about the article. On the other hand, news source metadata, which can be vital for the EBS context, is not directly accessible. This kind of metadata is produced through exogenous information or induced information about the news sources. We analysed both metadata to improve the data quality for the classification task. We divided this contribution into mainly two different classification tasks, i.e., 1) Metadata-Based News Article Classification and 2) Metadata-Based News Source Classification. The details of these tasks are as follows:

6.1 Metadata-Based News Article Classification

Metadata in news articles helps in organizing and managing news content within the context of EBS. It enables efficient search and retrieval of articles based on specific criteria, such as date, topic, or source. In our proposed approach, we extracted the above-mentioned metadata through web scrapping. As we will use this metadata as features for machine learning model.

Data preprocessing is an essential step in building machine learning models with text data. It involves cleaning and transforming the raw text data into a format that can be easily understood and processed by machine learning algorithms [71]. Some preprocessing techniques include lowercasing, removing stop words, stemming etc. For instance, we applied stop word removal on metadata textual features i.e., title, description by removing stop words from the text. Subsequently, we applied stemming and lemmatization techniques using *spaCy* on metadata features i.e., title, description to standardize the text. This text standardization converts the word into its base form by removing prefix, suffix or reduction of the word so that it could be easier to analyse by models. Additionally, we applied URL Tokenization on specific

metadata features such as 'URL' and 'Source' by converting the URL into valid tokens. These tokens of source and URL can be useful for machine learning model for the classification task. In order to tokenize, we established a regular expression to split the URL into words or tokens. For instance, the URL "http://www.sample.com/level1/index.html?id=1234" is split into valid words i.e., 'http', 'www', 'sample', 'com', 'level1', 'index', 'HTML', 'id', '1234'. At the end of this step, we have a set of features for model from the metadata attributes.

In next step, we performed experiments with several machine learning models i.e. Logistic Regression (LR), Support vector classifier (SVC) and Stochastic Gradient Descent (SGD) Classifier for the classification of relevant news articles. The idea is to see which article metadata features are important for the classification. Among the three models, SGD was performing slightly better among SVC and LR classifiers using metadata features. Stochastic Gradient Descent (SGD) Classifier is a linear classifier that is efficient and can handle sparse data well [72-73]. It is often used in text classification problems where the number of features is limited as compared to the size of the dataset. The goal of this approach is to classify relevant news articles in a more resource-efficient manner. The process pipeline for the approach for classification through metadata of news article is shown in Figure 3.

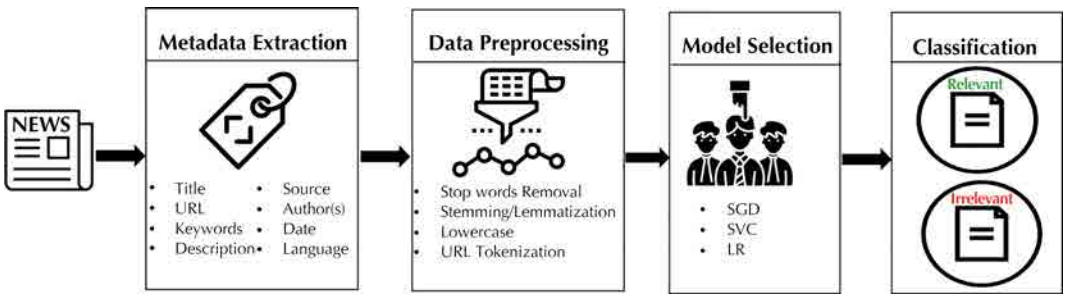

Figure 3. Workflow: Classification through Article Metadata.

6.2 Metadata-Based News Source Classification

In the second task, we identified news source metadata features for the source classification task. These source metadata include source category, e.g. specialized or generalized, geographical coverage, media bias and topic coverage. In the context of EBS, analysing news source metadata can be helpful in identifying authoritative and specialized sources [74] that are more likely to provide accurate and relevant information about specific events or topics [75]. The details of these source metadata features associated with the news sources are as follows:

1. **Source Specialization:** In the context of EBS systems, apart from government official sources, dedicated health sources mainly reports about public health, disease outbreaks and emergency preparedness. For instance, **Outbreak News Today**[①] is an online news source that reported the news

[①] https://outbreaknewstoday.com/

about various infectious diseases and their outbreaks. In Figure 4, we show different specialized online news sources that mainly report news about agriculture and poultry issues.

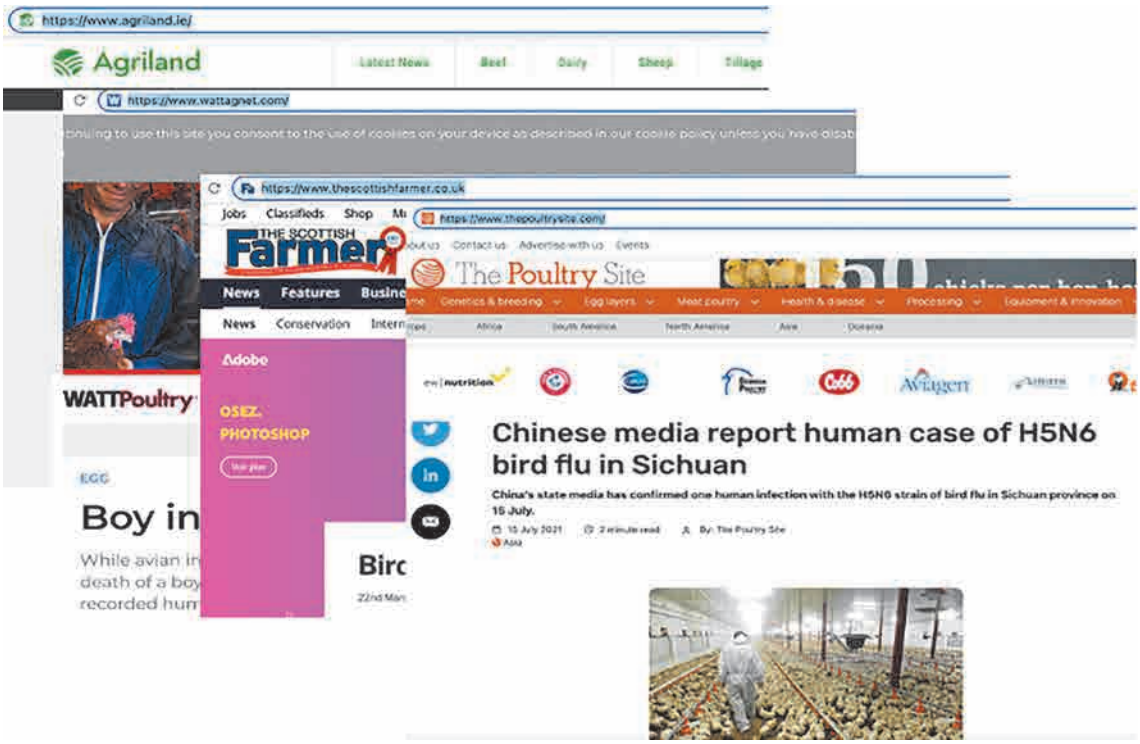

Figure 4. Specialized Sources of Poultry and Agriculture.

The automatic news source categorization is not a straight forward task. In order to achieve it, the first step is to extract the title and description information of the news source. The title and description of the news sources are extracted through web scrapping. Subsequently, *Google Translate API*[2] is utilized to check the language of the news source from the text of title and description. Afterwards, the text of the title and description is further translated using *Google Translate API* in case of non-English text of title and description. Google Translate API was selected as the preferred choice due to its renowned reputation for delivering accurate translations, distinguishing it from other freely accessible libraries. In the existing literature, clustering is one of the most famous text mining technique used for categorization of textual documents [76-77]. Because of its unsupervised nature, it is often preferred for categorizing text documents. K-means is a straightforward and easy-to-understand clustering algorithm, preferred for simpler and quick clustering task. Therefore, we applied k-means clustering technique using textual features in title and description for source categorization, i.e., Specialized and Generalized. For clustering, we used a

[2] https://py-googletrans.readthedocs.io/en/latest/

specific dictionary of the terms relevant in the context of avian-influenza disease. Figure 5 shows the flowchart to classify the news source into generalized and specialized category. The final output of the flow chart is source category (specialized/generalized).

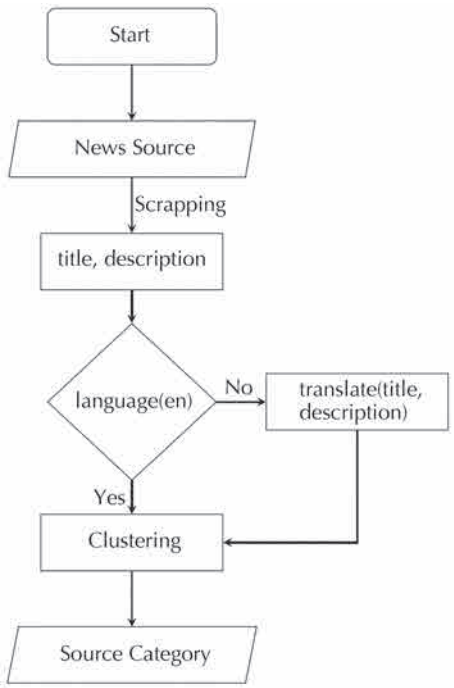

Figure 5. Flowchart of Source Categorization.

2. **Geographical Coverage:** Local news sources excel in providing immediate, detailed information about health events at the community level, emphasizing the local impact and response. On the other hand, international source provides more global perspective with a focus on comprehensive coverage and analysis of national and international health events. Therefore, the level of information and timeliness may vary in both cases. Due to the level of information and timeliness, it is important to take into account this aspect about news sources. Figure 6 shows geographical coverage of three different web sources. Examples of geographical coverage of MidiLibre (1) as *local* news source of France, Farmers (2) as *national* news source of the United Kingdom (UK) and CNN (3) as *international* news source.

   To our knowledge, there is no such method to automatically extract the geographical coverage of the news source. However, by analysing different news sources, we found some pattern to identify the geographical coverage of the news source. Geographical coverage is usually available in the menu of main page of the news source websites. The menu contains geographical references e.g., world, country names, region names, city names etc. By following these patterns, we developed our

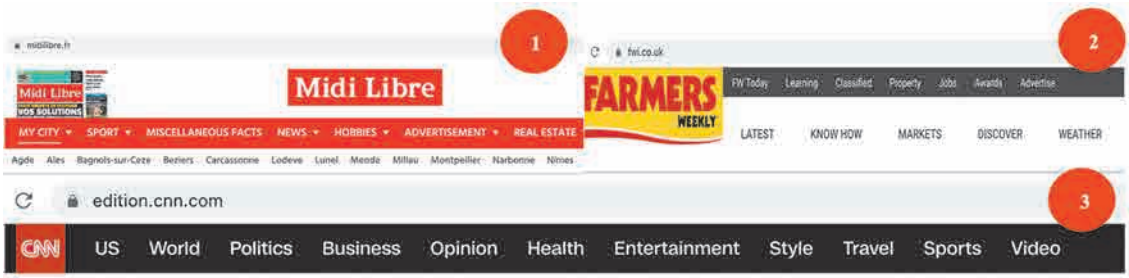

Figure 6. Geographical Coverage: Local, National and International News Source.

custom algorithm to extract this information using web scrapping and NLP techniques. The steps followed to extract the geographical coverage are as follows:

- Select the URL (home page) of the news source.
- Analyse the webpage of the news source, and extract the *locations* from the menu items of the webpage.
- If the *locations* are continents or countries, it is said to be an International news source.
- If the *locations* are cities, then find its province/region using geocoding API (geopy[3] python library).
- If the geocoded cities belong to a single region/province/state (they are often similar) can be a local news source. If it belongs to multiple regions of the same country, it can be a national news source.

3. **Media Bias:** Biased reporting may exaggerate the severity of the health situation or downplay it based on the agenda of the news source [78]. This can lead to public confusion and panic, resulting in difficulties in effective public health responses. In order to see the media bias, World Press Freedom Index (wpfi) is an assessment measure of press freedom and the level of media independence in countries around the world [79]. The aim is to assess media freedom, highlight countries where press freedom is restricted or violated, and the promotion of free and independent media. Figure 7 shows media freedom index examples of countries, e.g., Netherlands, Norway and Nigeria.
   The wpfi freedom index for the countries are automatically extracted through web scrapping from Reporters without border (RSF) website [80].
4. **Topic Coverage:** Topic specific news sources play significant roles in disease outbreak detection, monitoring, and response. For instance, agriculture related news sources often report on crop diseases or livestock illnesses that can serve as early warning signs of potential zoonotic diseases that can be transmitted to other humans and animals. Livestock-focused news sources report on diseases affecting animals, including those that may have zoonotic potential. For the experiments, we curated this information about the news sources from SimilarWeb[4] which is a website analytic tool. Experiments are performed to classify the news sources using the mentioned identified news source metadata features. The experiments performed with results are discussed in the subsequent section.

[3] https://pypi.org/project/geopy/
[4] https://www.similarweb.com/

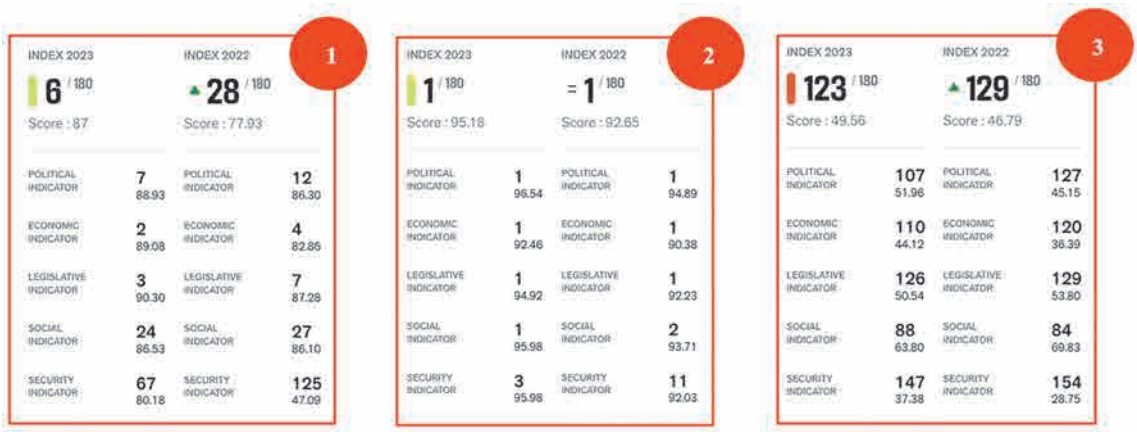

Figure 7. World Press Freedom Index of the Netherlands, Norway, Nigeria.

Data and Software Availability

The whole workflow in this chapter is divided into two main approaches, i.e., 1) Data driven approach for news classification, and 2) Metadata approach for news classification. The code, datasets and results are available at GitHub repository[5].

7. RESULTS

The results are obtained for the following approaches: 1) A data-driven score-based approach, 2) Metadata-based article classification, and 3) Metadata-based source classification. The specifics of these results are outlined below:

7.1 A data-driven score-based approach

The news classification with score-based approach is evaluated though precision, recall, and the F-Score measures [81-82]. The definitions of precision, recall, and the F-Score are as follows [83]:

$$Precision = \frac{Correctly\ \ Relevant\_News\_Articles\ Classified}{Total\ \ Relevant\_News\_Articles\ Classified} \tag{5}$$

$$Recall\ = \frac{Correctely\ \ Relevant\_News\_Articles\ Classified}{Total\ \ Relevant\_News\_Articles\ in\ Dataset} \tag{6}$$

$$F\ -Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{7}$$

[5] https://github.com/mehtab-alam/data_quality.git

Table 1 shows the precision, recall, and F-score for different components scores (MS, CS, E3S, DQS) used to classify news articles into relevant (outbreak-related news articles) and irrelevant (no outbreak event) categories (see Section 4.1). The results show varying performance across these scoring methods. The Metadata Score demonstrated moderate recall (0.7) and precision (0.44), with an F-Score of 0.54, suggesting a balanced but not highly effective performance. The Content Score, on the other hand, exhibited a better precision (0.71) but very low recall (0.1), resulting in a low F-Score of 0.14, indicating a significant problem with false negatives. Furthermore, the Epidemiological Entity Extraction Score (E3S) performed reasonably well, with a moderate F-Score of 0.53, reflecting a balanced performance in terms of precision (0.61) and recall (0.46). However, the DQS which is based on the average of all scores performed better than others in terms of F-Score of 0.8, combining exceptionally high precision (0.97) with a respectable recall (0.68).

Table 1. Results: Data Driven Approach.

Score Type	Precision	Recall	F-Score
MS	0.44	0.70	0.54
CS	0.71	0.10	0.14
E3S	0.61	0.46	0.53
DQS	0.97	0.68	0.80

These results highlight the critical importance of selecting an appropriate scoring mechanism when classifying news articles in disease outbreak detection. While high precision is valuable to minimize false positives, a balance with recall is essential to avoid overlooking relevant articles. In this context, the DQS stands out as a robust choice, offering both high precision and reasonable recall. However, the choice of scoring method should align with specific goals, and the trade-offs between precision and recall should be carefully considered based on the desired outcomes of the classification task. This approach is valuable because it allows us to assess how different attributes influence the quality of news articles. Furthermore, in the multidisciplinary MOOD project involving end-users, attaining this degree of explainability is more significant.

7.2 Metadata-Based News Article Classification

We performed the classification through individual metadata feature and the combination of all metadata features in order to see the important metadata features. Table 2 shows the results which contain news articles extracted from PADI-web from relevant class and irrelevant class. Table 2 provides the evaluation results of a system that has been trained to classify through metadata features as either 'Relevant' or 'Irrelevant' (see Section 4.1). The evaluation metrics used are precision, recall, and F-score. Overall, the system performs well, with the highest F-score being 0.96 for the "All" parameter, indicating that the system is able to make accurate and precise predictions.

Table 2. Results: Metadata-Based News Article Classification.

Metadata Attributes	Precision	Recall	F-Score
URL	0.86	0.87	0.86
Source	0.76	0.74	0.75
title	0.94	0.94	0.94
description	0.9	0.82	0.84
publish date	0.45	0.47	0.44
keywords	0.92	0.93	0.92
authors	0.73	0.61	0.59
language	0.31	0.5	0.38
All	0.97	0.95	0.96

The two best performing metadata features according to the table are 'Title' and 'Keywords'. 'Title' has the highest recall of 0.94 and an F-score of 0.94, meaning that the system is accurately classifying most of the metadata features as either 'Relevant' or 'Irrelevant'. 'Keywords' has a precision of 0.92, a recall of 0.93, and an F-score of 0.92, indicating that the system is making accurate predictions and finding most of the relevant results. The confusion matrix for the Table 2 results for the following metadata features 'Title', 'Keywords' and 'All' are shown in Figure 8.

7.3 Metadata-Based News Source Classification

Our main objective was to identify interesting features associated with the news sources. We evaluate our results at two levels, i.e., 1) feature Classification e.g. source category and geographical coverage, and 2) Source Classification. The evaluation of source categorization is done in supervised way. In the dataset (see Section 4.2), we manually annotated the source category of the news source to evaluate the automated source categorization by comparing *source category* with annotated category. The clustering results into two main categories, i.e., specialized, generalized. Clustering techniques produced better results in source categorization with precision of **0.88**, recall of **0.83** and F-score of **0.86**. This category is further used for the classification of news sources. We used this source metadata as a feature for the source classification task.

In the dataset (see Section 4.2), we manually annotated the geographical coverage of the news source to evaluate the automated geographical coverage. Table 3 shows the evaluation of the geographical coverage categorization of news sources, with recall of 0.37 for local class, 0.57 for national class and 0.87 for international class. The results show less recall for categorization of local and national sources. Therefore, the algorithm needs significant improvements for categorization of local and national sources.

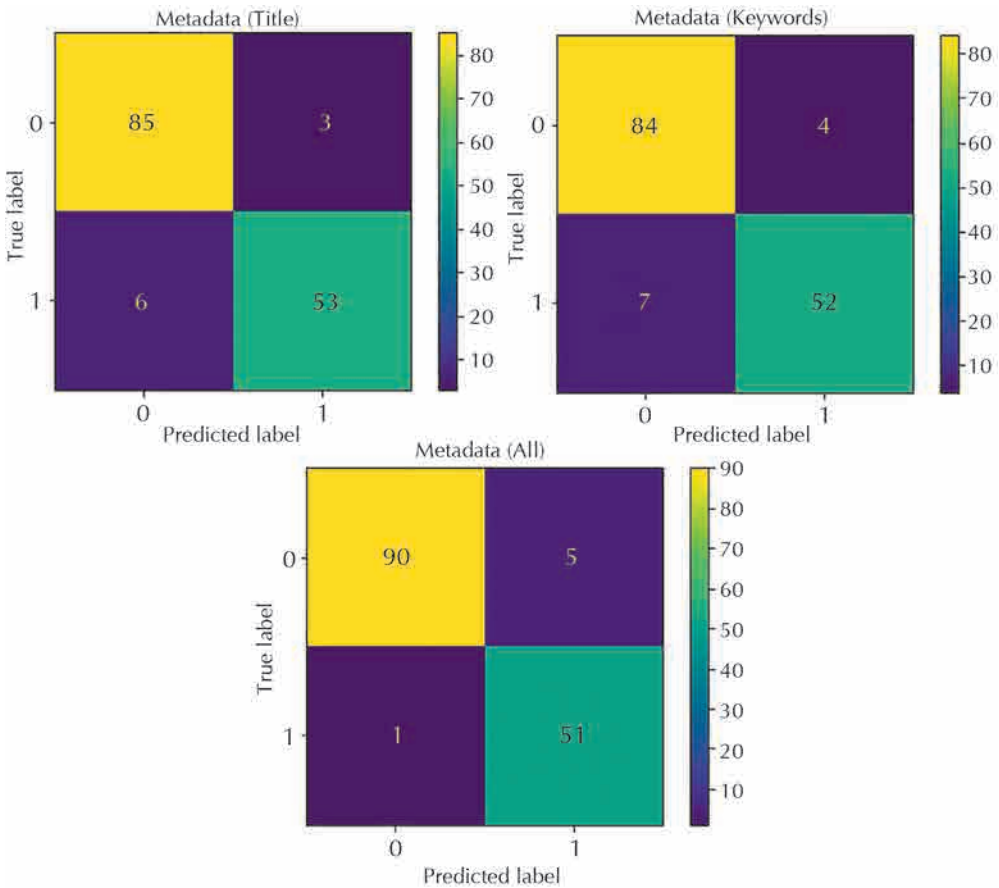

Figure 8. Confusion Matrices of Metadata-Based News Article Classification (Title, Keywords, All Features).

Table 3. Geographical Coverage Recall Measure.

Predicted Labels	True Labels			Recall
	Local	National	International	
Local	6	4	6	0.37
National	4	12	5	0.57
International	6	5	78	0.88
Recall				0.76

We produced preliminary results for the classification of news sources for the dataset discussed in Section 4.2. We selected the Random Forest model as the best choice for classifying sources based on numerical (wpfi) and categorical features (Specialization, Geographical Coverage and Topic Coverage).

Table 4 shows the results of classification of relevant and irrelevant news sources, with F-score of 0.85 for irrelevant class and 0.47 for relevant class.

Table 4. Results: Metadata-Based News Source Classification.

Features	Class	Precision	Recall	F-Score	Accuracy
Source Category Geographical Coverage Media Bias Topic Coverage	Relevant	0.74	1	0.85	**0.76**
	Irrelevant	**1**	**0.31**	0.47	

The most important features for the random forest model for classifying news sources were 'topic coverage' and 'Media Bias (wpfi)'. The results are not promising at this stage, but these features in addition to more source metadata features can be helpful for the quality quantification of news sources.

8. DISCUSSION

In the score-based approach presented in Section 5, we comprehensively considered contextual information attributes to ensure the relevance of news sources. These attributes encompass metadata, content, and epidemic-related aspects, including temporal, spatial, and epidemic information. The collective score derived from these weighted entities results in the contextual weight of each article. The Data Quality Score (DQS) plays a pivotal role in classifying news articles into two main categories: 'Relevant' and 'Irrelevant'. The research has contributed significantly by focusing on a dataset specifically dedicated to AI related articles. Nonetheless, certain limitations should be acknowledged. A primary limitation is the use of large, which may not fully represent the diversity and complexity of articles related to different diseases. To address this, future work should involve larger datasets to explore patterns and insights across various diseases comprehensively.

Metadata extraction holds a crucial position within the EBS System pipeline. The Metadata-Based News Article Classification discussed in Section 6.1 has successfully highlighted key metadata features, including URL, source, keywords, and article titles. These features significantly contribute to the accurate classification of relevant news articles. While these findings are promising, limitations persist, particularly related to validation on smaller datasets. Tokenizing the URL and source is critical for successful article classification within the EBS System. However, it is essential to recognize that not all URLs contain relevant keywords for classification, with many news sources employing URL patterns that comprise only an ID, lacking meaningful information for the classification model.

In addition to that, we incorporated vital attributes such as source category, geographical coverage, topic coverage, and media bias (wpfi) to classify relevant news sources. Categorizing news sources based on their specializations or focus areas allows the EBS to filter and prioritize information for specific disease-related events. For instance, sources specializing in poultry and agriculture news can receive

more weight when tracking AI disease outbreaks, while those focusing on finance and economics can be valuable for assessing diseases economic impacts. Additionally, understanding a news source coverage extent enables targeted monitoring of events in particular regions or countries, which is crucial for detecting and responding to geographically localized events. Moreover, the wpfi attribute into the metadata offers a quantifiable measure of a news source credibility and reputation within the media landscape.

9. CONCLUSION & PERSPECTIVES

In this work, we investigated the assessment of data quality of online news sources within the context of the EBS pipeline. We mainly proposed the three approaches to assess the quality of news sources. The first approach, named as, "a data-driven score-based approach to assess the quality of news articles" classify relevant news articles in the context of EBS. This method ensure the explainability aspect of attributes that makes it understandable for the end-users perspectives. The limitation of this work is its use of a small dataset. To address this limitation, future work should involve larger datasets to explore patterns and insights across various diseases comprehensively.

The second approach, named as, "Metadata-based news article classification" is proposed to classify relevant news articles in the context of EBS. In addition to the first approach, the benefit of this approach is to classify relevant news articles with limited features (metadata) without using content features through machine learning models. The finding was that metadata features like 'title' and 'keywords' are more important for the classification task. The limitation of this work was evaluation with the same dataset as discussed in the first approach.

The third approach, named as, "Metadata-based news source classification" is used for the categorization of news sources. This approach can help in prioritization of news sources in the context of EBS. However, we are still investigating other attributes to enhance this approach. For instance, 'Timely Reporting' is also a potential avenue for investigation that will help in quantification of recent events reported by news source. A source frequently reporting on outdated events would be assigned a lower weight, considering the temporal relevance of its content. This approach enhances the accuracy and timeliness of assessments. The significance of this study is that it can be a part of EBS pipeline in order to filter the noisy data to have more efficacy of automated EBS systems. In our case, we are more focused in to integrate it in PADI-web (https://padi-web.cirad.fr/en/) an EBS system.

Currently, the different approaches were applied to the news articles dataset for the disease case study of Avian Influenza. However, our future plans include testing our approaches to generalize and evaluate them with other disease datasets. For instance, we aim to consider plant disease data as illustrated in the study by EFSA[©]. Additionally, we have the PADI-web for Plant Health [84], which provide concrete perspectives and enhance our methodologies further. The ultimate objective is to seamlessly integrate these quality attributes into the PADI-web system, associating quality labels with both news sources

[©] https://efsa.onlinelibrary.wiley.com/doi/epdf/10.2903/sp.efsa.2016.EN-1118

and news articles. Enriched metadata features will further enhance the classification task within the pipeline, providing valuable insights and explanations for end-users. Furthermore, the research will be extended to include a two-level classification of news articles. In the case of relevant articles, it will be further classified into specific contextual classes, such as outbreak declarations, risk concerns, disease transmission, preventive, and control measures.

ACKNOWLEDGMENTS

This study was partially funded by EU grant 874850 MOOD and is catalogued as MOOD106. The contents of this publication are the sole responsibility of the authors and do not necessarily reflect the views of the European Commission. This work has also been funded by DGAL and the French National Research Agency under the Investments for the Future Program: ANR-16-CONV-0004 (#DigitAg).

AUTHOR CONTRIBUTION STATEMENT

Mehtab Alam Syed: Methodology, Software, Data curation, Writing-Review & Editing. **Elena Arsevska**: Data curation, Review & Editing. **Mathieu Roche**: Data curation, Guideline, Methodology, Review & Editing. **Maguelonne Teisseire**: Data curation, Methodology, Guideline, Review & Editing.

REFERENCES

[1] Mira Kim, Kyunghee Chae, Seungwoo Lee, Hong-Jun Jang, and Sukil Kim. Automated classification of online sources for infectious disease occurrences using machine-learning–based natural language processing approaches. *International Journal of Environmental Research and Public Health*, 17(24):9467, 2020.

[2] EE Rees, V Ng, P Gachon, A Mawudeku, D McKenney, J Pedlar, D Yemshanov, J Parmely, and J Knox. Early detection and prediction of infectious disease outbreaks. *CCDR*, 45:5, 2019.

[3] WHO. A guide to establishing event–based surveillance. *World Health Organization*, 2008.

[4] Silvia Runge-Ranzinger, Olaf Horstick, Michael Marx, and Axel Kroeger. What does dengue disease surveillance contribute to predicting and detecting outbreaks and describing trends? *Tropical Medicine & International Health*, 13(8):1022–1041, 2008.

[5] WHO. Who announces covid–19 outbreak a pandemic, 2020.

[6] Auss Abbood, Alexander Ullrich, Rüdiger Busche, and Stéphane Ghozzi. Eventepi—a natural language processing framework for event–based surveillance. *PLoS computational biology*, 16(11):e1008277, 2020.

[7] S Arunmozhi Balajee, Stephanie J Salyer, Blanche Greene-Cramer, Mahmoud Sadek, and Anthony W Mounts. The practice of event-based surveillance: concept and methods. *Global Security: Health, Science and Policy*, 6(1):1–9, 2021.

[8] Clark C. Freifeld, Kenneth D. Mandl, Ben Y. Reis, and John S. Brownstein. 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, 2008.

[9] Victor L Yu and Lawrence C Madoff. Promed-mail: an early warning system for emerging diseases. *Clinical infectious diseases*, 39(2):227–232, 2004.

[10] Michael Blench. Global public health intelligence network (GPHIN). In Proceedings of the 8th Conference of the Association for Machine Translation in the Americas: Government and Commercial Uses of MT, pages 299–303, Waikiki, USA, 2008. Association for Machine Translation in the Americas.

[11] Philip Abdelmalik, Emilie Peron, Johannes Schnitzler, Julie Fontaine, Eva Elfenkampera, and Philippe Barbozaa. The epidemic intelligence from open sources initiative: a collaboration to harmonize and standardize early detection and epidemic intelligence among public health organizations/l'initiative «epidemic intelligence from open sources»: une collaboration visant a harmoniser et a standardiser les procedures de detection precoce et de renseignement epidemiologique entre les organisations de sante publique. *Weekly Epidemiological Record*, 93(20):267–270, 2018.

[12] Kenrick D Cato, Bevin Cohen, and Elaine Larson. Data elements and validation methods used for electronic surveillance of health care-associated infections: A systematic review. *American journal of infection control*, 43(6):600–605, 2015.

[13] Michael Y Lin, Bala Hota, Yosef M Khan, Keith F Woeltje, Tara B Borlawsky, Joshua A Doherty, Kurt B Stevenson, Robert A Weinstein, William E Trick, CDC Prevention Epicenter Program, et al. Quality of traditional surveillance for public reporting of nosocomial bloodstream infection rates. *JAMA*, 304(18):2035–2041, 2010.

[14] Jens P Linge, Ralf Steinberger, TP Weber, Roman Yangarber, Erik van der Goot, DH Al Khudhairy, and NI Stilianakis. Internet surveillance systems for early alerting of health threats. *Eurosurveillance*, 14(13):19162, 2009.

[15] Elena Arsevska, Sarah Valentin, Julien Rabatel, Jocelyn De Goër de Hervé, Sylvain Falala, Renaud Lancelot, and Mathieu Roche. Web monitoring of emerging animal infectious diseases integrated in the french animal health epidemic intelligence system. *PLoS One*, 13(8):e0199960, 2018.

[16] Malwina Carrion and Lawrence C. Madoff. ProMED-mail: 22 years of digital surveillance of emerging infectious diseases. *International Health*, 9(3):177–183, 06 2017.

[17] Jens P Linge, Ralf Steinberger, Flavio Fuart, Stefano Bucci, Jenya Belyaeva, Monica Gemo, Delilah Al-Khudhairy, Roman Yangarber, and Erik van der Goot. Medisys: medical information system. In *Advanced ICTs for disaster management and threat detection: Collaborative and distributed frameworks*, pages 131–142. IGI Global, 2010.

[18] Sarah Valentin, Elena Arsevska, Julien Rabatel, Sylvain Falala, Alizé Mercier, Renaud Lancelot, and Mathieu Roche. Padi-web 3.0: A new framework for extracting and disseminating fine–grained information from the news for animal disease surveillance. *One Health*, 13:100357, 2021.

[19] Michael A. Cacciatore. Misinformation and public opinion of science and health: Approaches, findings, and future directions. *Proceedings of the National Academy of Sciences*, 118(15):e1912437117, 2021.

[20] Heidi Abbas, Mohamed Mostafa Tahoun, Ahmed Taha Aboushady, Abdelrahman Khalifa, Aura Corpuz, and Pierre Nabeth. Usage of social media in epidemic intelligence activities in the who, regional office for the eastern mediterranean. *BMJ Global Health*, 7(Suppl 4):e008759, 2022.

[21] Cheng Zhou, Haoxin Xiu, Yuqiu Wang, and Xinyao Yu. Characterizing the dissemination of misinformation on social media in health emergencies: An empirical study based on covid–19. *Information Processing & Management*, 58(4):102554, 2021.

[22] Zach Bastick. Would you notice if fake news changed your behavior? an experiment on the unconscious effects of disinformation. *Computers in human behavior*, 116:106633, 2021.

[23] Kumanan Wilson and John S Brownstein. Early detection of disease outbreaks using the internet. *Cmaj*, 180(8):829–831, 2009.

[24] Richard Y Wang and Diane M Strong. Beyond accuracy: What data quality means to data consumers. *Journal of management information systems*, 12(4):5–33, 1996.

[25] Carlo Batini, Monica Scannapieco, et al. Data and information quality. *Cham, Switzerland: Springer International Publishing. Google Scholar*, 43, 2016.

[26] Adam T Craig, Mike Kama, Marcus Samo, Saine Vaai, Jane Matanaicake, Cynthia Joshua, Anthony Kolbe, David N Durrheim, Beverley J Paterson, Viema Biaukula, et al. Early warning epidemic surveillance in the pacific island nations: an evaluation of the pacific syndromic surveillance system. *Tropical Medicine & International Health*, 21(7):917–927, 2016.

[27] Nadia K Conroy, Victoria L Rubin, and Yimin Chen. Automatic deception detection: Methods for finding fake news. *Proceedings of the association for information science and technology*, 52(1):1–4, 2015.

[28] Jane Mandalios. Radar: An approach for helping students evaluate internet sources. *Journal of information science*, 39(4):470–478, 2013.

[29] Xiaolan Zhu and Susan Gauch. Incorporating quality metrics in centralized/distributed information retrieval on the world wide web. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 288–295, 2000.

[30] Yoshiko Nozato. Credibility of online newspapers. Convención Anual de la Association for Education in Journalism and Mass Communication. Washington, DC Disponible en: http://citeseerx.ist.psu.edu/viewdoc/summary, 2002.

[31] Philipp Bachmann, Mark Eisenegger, and Diana Ingenhoff. Defining and measuring news media quality: Comparing the content perspective and the audience perspective. *The International Journal of Press/Politics*, page 1940161221999666, 2021.

[32] Md Momen Bhuiyan, Amy X Zhang, Connie Moon Sehat, and Tanushree Mitra. Investigating differences in crowdsourced news credibility assessment: Raters, tasks, and expert criteria. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–26, 2020.

[33] Laura Slaughter, Alla Keselman, Andre Kushniruk, and Vimla L Patel. A framework for capturing the interactions between laypersons' understanding of disease, information gathering behaviors, and actions taken during an epidemic. *Journal of biomedical informatics*, 38(4):298–313, 2005.

[34] Giandomenico Di Domenico, Jason Sit, Alessio Ishizaka, and Daniel Nunan. Fake news, social media and marketing: A systematic review. *Journal of Business Research*, 124:329–341, 2021.

[35] Marwa Essam and Tamer Elsayed. Why is that a background article: A qualitative analysis of relevance for news background linking. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 2009–2012, 2020.

[36] Mohamed K Elhadad, Kin Fun Li, and Fayez Gebali. A novel approach for selecting hybrid features from online news textual metadata for fake news detection. In *International Conference on P2P, Parallel, Grid, Cloud and Internet Computing*, pages 914–925. Springer, 2019.

[37] Md Rafiqul Islam, Shaowu Liu, Xianzhi Wang, and Guandong Xu. Deep learning for misinformation detection on online social networks: a survey and new perspectives. *Social Network Analysis and Mining*, 10(1):1–20, 2020.

[38] Yang Hu, Mingjing Li, Zhiwei Li, and Wei-ying Ma. Discovering authoritative news sources and top news stories. In *Asia Information Retrieval Symposium*, pages 230–243. Springer, 2006.

[39] Junting Ye and Steven Skiena. Mediarank: computational ranking of online news sources. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2469–2477, 2019.

[40] Charlotte Rudnik, Thibault Ehrhart, Olivier Ferret, Denis Teyssou, Raphael Troncy, and Xavier Tannier. Searching news articles using an event knowledge graph leveraged by wikidata. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, page 1232–1239, New York, NY, USA, 2019. Association for Computing Machinery.

[41] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36, 2017.

[42] Verónica Pérez–Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. Automatic detection of fake news. *arXiv preprint arXiv:1708.07104*, 2017.

[43] Sherry L Vellucci. Metadata. Annual Review of Information Science and Technology (ARIST), 33:187–222, 1998.

[44] Jenn Riley. Understanding metadata. Washington DC, United States: National Information Standards Organization (http://www.niso.org/publications/press/UnderstandingMetadata.pdf), 23:7–10, 2017.

[45] Tessel Bogaard, Laura Hollink, Jan Wielemaker, Jacco van Ossenbruggen, and Lynda Hardman. Metadata categorization for identifying search patterns in a digital library. *Journal of Documentation*, 75(2):270–286, 2019.

[46] Muzammil Khan, Arif Ur Rahman, M Daud Awan, and Syed Mehtab Alam. Normalizing digital news-stories for preservation. In *2016 Eleventh International Conference on Digital Information Management (ICDIM)*, pages 85–90. IEEE, 2016.

[47] Malte Ostendorff, Peter Bourgonje, Maria Berger, Julian Moreno–Schneider, Georg Rehm, and Bela Gipp. Enriching bert with knowledge graph embeddings for document classification. *arXiv preprint arXiv:1909.08402*, 2019.

[48] Gaël Lejeune, Romain Brixtel, Antoine Doucet, and Nadine Lucas. Multilingual event extraction for epidemic detection. *Artificial intelligence in medicine*, 65(2):131–143, 2015.

[49] Livio Cascone, Pietro Ducange, and Francesco Marcelloni. Exploiting online newspaper articles metadata for profiling city areas. In *International Conference on Intelligent Data Engineering and Automated Learning*, pages 203–215. Springer, 2019.

[50] Chenchen Ye, Linhai Zhang, Yulan He, Deyu Zhou, and Jie Wu. Beyond text: Incorporating metadata and label structure for multi-label document classification using heterogeneous graphs. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3162–3171, 2021.

[51] Naseer Ahmed Sajid, Munir Ahmad, Atta-ur Rahman, Gohar Zaman, Mohammed Salih Ahmed, Nehad Ibrahim, Mohammed Imran B Ahmed, Gomathi Krishnasamy, Reem Alzaher, Mariam Alkharraa, et al. A novel metadata based multi–label document classification technique. *Computer Systems Science & Engineering*, 46(2), 2023.

[52] Stephen Mutuvi, Emanuela Boros, Antoine Doucet, Gaël Lejeune, Adam Jatowt, and Moses Odeo. Multilingual epidemiological text classification: a comparative study. In *COLING, International Conference on Computational Linguistics*, pages 6172–6183, 2020.

[53] Jarutas Pattanaphanchai, Kieron O'Hara, and Wendy Hall. Trustworthiness criteria for supporting users to assess the credibility of web information. In *Proceedings of the 22nd international conference on world wide web*, pages 1123–1130, 2013.

[54] Piotr Przybyła and Axel J Soto. When classification accuracy is not enough: Explaining news credibility assessment. *Information Processing & Management*, 58(5):102653, 2021.

[55] Flávia Milhorance and J Singer. Media trust and use among urban news consumers in brazil. *Ethical Space: the international journal of communication ethics*, 15(3/4):56–65, 2018.

[56] Reijo Savolainen. Media credibility and cognitive authority. the case of seeking orienting information. *Information Research: An International Electronic Journal*, 12(3):n3, 2007.

[57] Xinyi Zhou and Reza Zafarani. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, 53(5):1–40, 2020.

[58] Ramy Baly, Mitra Mohtarami, James Glass, Lluís Màrquez, Alessandro Moschitti, and Preslav Nakov. Integrating stance detection and fact checking in a unified corpus. *arXiv preprint arXiv:1804.08012*, 2018.

[59] Oscar Alomar, Assumpció Batlle, Josep Maria Brunetti, Roberto García, Rosa Gil, Toni Granollers, Sara Jiménez, Amparo Laviña, Carme Reverté, Jordi Riudavets, et al. Development and testing of the media monitoring tool med is ys for the monitoring, early identification and reporting of existing and emerging plant health threats. *EFSA Supporting Publications*, 13(12):1118E, 2016.

[60] Reza Vaziri and Mehran Mohsenzadeh. A questionnaire–based data quality methodology. *International Journal of Database Management Systems*, 4(2):55, 2012.

[61] Lois Mai Chan, Eric Childress, Rebecca Dean, Edward T O'neill, and Diane Vizine-Goetz. A faceted approach to subject data in the dublin core metadata record. *Journal of Internet Cataloging*, 4(1-2):35–47, 2001.

[62] David Westerman, Patric R Spence, and Brandon Van Der Heide. Social media as information source: Recency of updates and credibility of information. *Journal of computer-mediated communication*, 19(2):171–183, 2014.

[63] Yuli Vasiliev. Natural Language Processing with Python and SpaCy: A Practical Introduction. No Starch Press, 2020.

[64] Nastaran Jafarpour, Masoumeh Izadi, Doina Precup, and David L Buckeridge. Quantifying the determinants of outbreak detection performance through simulation and machine learning. *Journal of biomedical informatics*, 53:180–187, 2015.

[65] Ken P Kleinman and Allyson M Abrams. Assessing surveillance using sensitivity, specificity and timeliness. *Statistical methods in medical research*, 15(5):445–464, 2006.

[66] World Health Organization et al. Early detection, assessment and response to acute public health events: implementation of early warning and response with a focus on event–based surveillance: interim version. Technical report, World Health Organization, 2014.

[67] Jochen L Leidner and Michael D Lieberman. Detecting geographical references in the form of place names and associated spatial natural language. *Sigspatial Special*, 3(2):5–11, 2011.

[68] James Pustejovsky, José M Castano, Robert Ingria, Roser Sauri, Robert J Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R Radev. Timeml: Robust specification of event and temporal expressions in text. *New directions in question answering*, 3:28–34, 2003.

[69] Angel X Chang and Christopher D Manning. Sutime: A library for recognizing and normalizing time expressions. In *LREC*, volume 3735, page 3740, 2012.

[70] Steffen Lohmann, Florian Heimerl, Fabian Bopp, Michael Burch, and Thomas Ertl. Concentri cloud: Word cloud visualization for multiple text documents. In *2015 19th International Conference on Information Visualisation*, pages 114–120. IEEE, 2015.

[71] Kiran Maharana, Surajit Mondal, and Bhushankumar Nemade. A review: Data pre–processing and data augmentation techniques. *Global Transitions Proceedings*, 2022.

[72] Shilpa Gite, Shruti Patil, Deepak Dharrao, Madhuri Yadav, Sneha Basak, Arundarasi Rajendran, and Ketan Kotecha. Textual feature extraction using ant colony optimization for hate speech classification. *Big data and cognitive computing*, 7(1):45, 2023.

[73] Agung B Prasetijo, R Rizal Isnanto, Dania Eridani, Yosua Alvin Adi Soetrisno, Muhammad Arfan, and Aghus Sofwan. Hoax detection system on indonesian news sites based on text classification using svm and sgd. In *2017 4th international conference on information technology, computer, and electrical engineering (ICITACEE)*, pages 45–49. IEEE, 2017.

[74] Michael A Gisondi, Rachel Barber, Jemery Samuel Faust, Ali Raja, Matthew C Strehlow, Lauren M Westafer, and Michael Gottlieb. A deadly infodemic: social media and the power of covid–19 misinformation, 2022.

[75] Shuai Zhang, Feicheng Ma, Yunmei Liu, and Wenjing Pian. Identifying features of health misinformation on social media sites: an exploratory analysis. *Library Hi Tech*, 40(5):1384–1401, 2022.

[76] Sayali Sunil Tandel, Abhishek Jamadar, and Siddharth Dudugu. A survey on text mining techniques. In *2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS)*, pages 1022–1026. IEEE, 2019.

[77] Karwan Jacksi, Rowaida Kh Ibrahim, Subhi RM Zeebaree, Rizgar R Zebari, and Mohammed AM Sadeeq. Clustering documents based on semantic similarity using hac and k-mean algorithms. In *2020 International Conference on Advanced Science and Engineering (ICOASE)*, pages 205–210. IEEE, 2020.

[78] Zhan Xu. Examining us newspapers' partisan bias in covid-19 news using computational methods. *Communication Studies*, 74(1):78–96, 2023.

[79] Edina Berlinger, Judit Lilla Keresztúri, Ágnes Lublóy, and Zsuzsanna Vőneki Tamásné. Press freedom and operational losses: The monitoring role of the media. *Journal of International Financial Markets, Institutions and Money*, 77:101496, 2022.

[80] Reporters without borders rsf, 2012.

[81] Kai Hakala and Sampo Pyysalo. Biomedical named entity recognition with multilingual bert. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, pages 56–61, 2019.

[82] Philip Resnik and Jimmy Lin. 11 evaluation of nlp systems. *The handbook of computational linguistics and natural language processing*, 57, 2010.

[83] Cyril Goutte and Eric Gaussier. A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In *European conference on information retrieval*, pages 345–359. Springer, 2005.

[84] Mathieu Roche, Julien Rabatel, Carlène Trevennec, and Isabelle Pieretti. Padi-web for plant health surveillance. In Shareeful Islam and Arnon Sturm, editors, *Intelligent Information Systems*, pages 148–156, Cham, 2024. Springer Nature Switzerland.