A data-driven score model to assess online news articles in event-based surveillance system

Syed Mehtab Alam, Elena Arsevska, Mathieu Roche, Maguelonne Teisseire

A data-driven score model to assess online news articles in EBS

Mehtab Alam Syed[1,4], Elena Arsevska[2,5], Mathieu Roche[1,4], Maguelonne Teisseire[3,4]
[1]CIRAD, UMR TETIS, F-34398 Montpellier, France
[2]CIRAD, UMR ASTRE, F-34398 Montpellier, France
[3]INRAE, UMR TETIS, Montpellier
[4]TETIS, Univ Montpellier, AgroParisTech, CIRAD, CNRS, INRAE, Montpellier, France
[5]ASTRE, Univ Montpellier, CIRAD, INRAE, Montpellier, France
{mehtab-alam.syed,elena.arsevska,mathieu.roche}@cirad.fr,
maguelonne.teisseire@inrae.fr

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. However, these methods only rely on the extraction of spatio-temporal information for representing health events. To address this research gap, a score-based technique is proposed to quantify the data quality of online news articles through three assessment measures: 1) news article metadata, 2) content analysis, and 3) epidemiological entity extraction with NLP to weight the contextual information. The results are calculated using classification metrics with two evaluation approaches: 1) a strict approach and 2) a flexible approach. The obtained results show significant enhancement in the data quality by filtering irrelevant news, which can potentially reduce false alert generation in EBS systems.

Keywords: Text Mining, Natural Language Processing, Data Quality

1 Introduction

Outbreaks of infectious diseases pose serious threats to public health and safety (Kim et al., 2020). Infectious disease outbreaks affect not only public health but also the national and international economy and global awareness (Rees et al., 2019). It is important to implement public health surveillance methods to recognize potential infectious disease outbreaks and to minimize their associated devastating effects on society. In the existing literature (Organization and others, 2008; Rees et al., 2019), there are two main types of public health surveillance strategies: 1) event-based surveillance (EBS) and 2) indicator-based surveillance (IBS). Both of these surveillance strategies complement one another in terms of benefits due to their unique data collection, monitoring, assessment, and data interpretation processes (Organization and others, 2008) and are treated as fundamental in constructing a comprehensive surveillance system (Balajee et al., 2021). This research focuses on EBS, whereas IBS is outside the scope of this article.

EBS is the organized process of detecting and reporting information (i.e., represented as events) to healthcare authorities by rapidly capturing information from different unstructured data sources (Balajee et al., 2021). It enables concerned authorities to be better prepared for endemic and pandemic disease outbreaks by functioning as a key component of an effective early warning system (Organization and others, 2008; Balajee et al., 2021). For information acquisition, online information sources (e.g., news articles, blogs, rumors, social media (such as Twitter, etc.), and other ad hoc reports, etc.) have gained great attention in implementing "web-based" or "internet-based" EBS systems (Valentin, 2020) compared to traditional data collection methods, which are labor intensive (Cato et al., 2015) and restricted by inter-observer variability (Lin et al., 2010). The working efficiency of web-based EBS systems in terms of detecting true outbreaks to protect public health against the spread of infectious diseases depends on the quality of the information collected from online news sources (Valentin, 2020). As online news information is diverse and collected from heterogeneous online data sources, it is crucial to

verify this unstructured information to avoid misinformation (i.e., a piece of information that is false or having no scientific evidence) (Zhou et al., 2021) and disinformation (i.e., intentionally generated false information) (Bastick, 2021) that can pose serious threats to public health.

Unstructured news information sources are in a common textual format (Valentin, 2020; Carneiro and Mylonakis, 2009). Large amounts of textual information from diverse information sources can overwhelm web-based EBS systems. Existing EBS-based methods exploiting different text mining techniques transform online textual data into a computer-readable format to enable the extraction of relevant news information from existing textual sources per human needs (Arsevska et al., 2016; Valentin, 2020). However, this method not only relies on the extraction of spatiotemporal information (i.e., when-where questions) for representing health events but also does not adequately evaluate the accuracy of these executed mapping processes for extracting the correct information for generating true health alerts (Ganser, 2020).

The work presented in this paper aims to address these research gaps by 1) including data quality attributes based on metadata and news content through the extraction of epidemiological information (in addition to the extraction of spatiotemporal data) to identify relevant information. This epidemiological information is extracted using a pattern-based text mining approach. This concerns the epidemiological concepts and terms related to infectious diseases and particularly avian influenza. 2) Then, news sources are labeled into two different groups of news articles, relevant (outbreak-related) and irrelevant, based on the data quality scores. Consequently, the proposed method prioritizes outbreak-related news by discarding irrelevant news in real time to minimize false health alerts in EBS systems.

The paper is structured as follows: Section 2 describes the state-of-the-art literature related to quality attributes and the evaluation of online news sources. Section 3 presents the proposed methodology. Section 4 presents the results of experiments, and Section 5 discusses the proposed work. Finally, Section 6 presents the conclusion and outlines future work.

2 State of the art

Research on data quality, which is crucial for evaluating online news sources and constructing EBS systems, began in the 1990s. Wang and Strong (Wang and Strong, 1996) 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 (Batini et al., 2016). 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 (Mandalios, 2013) 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 (Zhu and Gauch, 2000) 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 (Nozato, 2002) stated that the timeliness, depth, reputation, and accuracy of online sources are the most important data quality dimensions. Another study (Bachmann et al., 2021) used the quality attributes of the respondents and general perception of the news sources for news classification. Moreover, another study (Bhuiyan et al., 2020) investigated news credibility assessments by comparing crowds and expert opinions to understand the differentiation in the rating of the source.

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 (Cato et al., 2015) based on information retrieval, machine learning, deep learning and knowledge representation graphs for assessing the relevance of news sources. For example, Essam and Elsayed (Essam and Elsayed, 2020) defined a specialized information retrieval technique by assessing the topics and subtopics of the news to identify highly relevant background articles. Elhadad et al. (Elhadad et al., 2019) 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 (Islam et al., 2020) 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. (Hu et al., 2006) analyzed 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 also limited to small news corpora. To address this limitation, a system named MediaRank was designed (Ye and Skiena, 2019) to incorporate large datasets for measuring the quality of news sources by a mix of computational signals reflecting peer reputation, reporting bias, bottomline pressure, and popularity. A study employing the application of knowledge graphs by Rudnik et al. (Rudnik et al., 2019) implemented a method using a Wikidata knowledge base for generating the semantic annotation of news articles to filter relevant news articles.

As stated above, substantial efforts have been made in news article classification using different data quality dimensions and methods. However, these studies have been centered on generalized news article classification problems. For designing a web-based EBS system, a domain-oriented news article classification approach is needed to filter relevant news articles. In the literature, there exist numerous EBS systems that not only perform manual curation of spatiotemporal entities but also map them with domain-specific (thematic) entities (Arsevska et al., 2016). Therefore, the accuracy of these event-related extracted information from the processed mapping process is inappropriate for generating true alerts (Ganser, 2020). For instance, the existing HealthMap system provides the latitude and longitude coordinates linked to each event. However, the accuracy of the extracted geographical features has never been evaluated. Moreover, various other state-of-the-art systems, including HealthMap (Ganser, 2020; Valentin, 2020), use the news publication date to determine the occurrence date of the event. However, the extraction of temporal information directly from news articles has not yet been sufficiently explored and validated. To fill this research gap, a study proposed by Alomar et al. (Alomar et al., 2016) 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. Alomar et al. (Alomar et al., 2016) discussed a direct method (i.e., identification, review and evaluation of known sources to find relevant information sources) and an indirect method that assesses quality attributes of news content and metadata. Using the identified attributes by Alomar et al. (Alomar et al., 2016), this work presents its extension by presenting the implementation and automatic extraction of attributes required for news article classification problems. To extend the baseline approach identified by Alomar et al. (Alomar et al., 2016) for extracting relevant news articles, this work reports a 3-step process of information extraction that uses state-of-the-art text mining techniques. The 1^{st} step involves the automatic extraction of the metadata information (i.e., publisher, subject, description, type, source, language, rights, and date) of the news article. This information is used to determine the trustworthiness of the article. The 2^{nd} step consists of evaluating the news article's content in terms of accessibility, relevance, accuracy, clarity, timeliness, and reputation parameters. Last, the 3^{rd} step, which is the main contribution of this work, introduces an epidemiological entity extraction approach (E3A) to weight the epidemiological context.

3 Proposed Work

Online information sources are the major information providers for event-based surveillance (EBS) systems. Therefore, the verification of these heterogeneous data sources is the primary concern of event-based systems. Thus, it is important to rate the data quality of online news sources by data quality scores before considering this news source information as part of an EBS system. In the proposed work, the data quality of online news sources is computed from three principal components: 1) metadata, 2) news content and 3) an epidemiological entity extraction approach (E3A). The main contribution of the E3A approach is introduced to improve the results by filtering out relevant news articles. The process workflow with all components is shown in Figure 1.

3.1 Data Quality Measures (DQM)

Data quality measures (DQMs) are metrics used for ranking high-quality and low-quality elements to filter reliable news sources from the perspective of relevance, accuracy and reputation (Vaziri

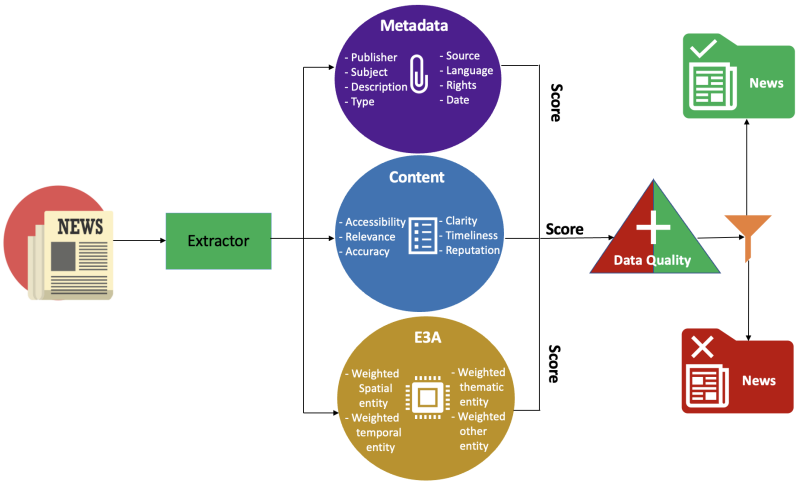

Figure 1: Process workflow

and Mohsenzadeh, 2012). There are many criteria used to compute the data quality of online news sources, i.e., metadata attributes and content analysis by attributes (Alomar et al., 2016). The proposed approach allows us to measure the metadata score (MS) from metadata components, the content score (CS) by the content component and the epidemiological entity extraction score (E3S) calculated with weighted named entities (i.e., spatiotemporal entities with epidemiological entities). The E3S is computed from weighted named entities using the E3A. These components are discussed in the subsequent sections.

3.1.1 Metadata

Metadata are the core components of data used to establish quick information (Chan et al., 2001). Therefore, metadata are defined and used to find relevant results quickly in search engines from millions of online sources (Chan et al., 2001). There are some tags defined by each reliable online source that are analyzed and fetched by search engines to provide ranked results.

There are different metadata attributes that are considered in the assessment of metadata from news sources (Alomar et al., 2016). However, some of the metadata attributes are not frequently available in the news sources and are ultimately discarded in the approach. We thus adopt eight attributes shortlisted as metadata that are considered relevant and available in most news articles. The following are the metadata attributes with the associated definitions:

- **Publisher**: Entity responsible for publishing news sources, e.g., Japan times, The Times of India
- **Subject**: Title of the news article, e.g., U.S. detected its first case of virulent bird flu in commercial poultry in 2017.
- **Description**: Short description of the news article
- **Type**: Type/nature of the news article, e.g., article
- **Source**: URL resource of the news article
- **Language**: Language of the news article, e.g., en_GB, en_US
- **Right**: Authority to publish the news source
- **Date**: Date of publication of news article. However, in some cases, the modification date of news articles is also available. The date of publication is the date at which the news article is published, and the modification date is the last updated date of the news article.

Each metadata attribute is extracted from the online news article. The MS is computed by the following formula (Alomar et al., 2016):

$$MS = \sum_{n=1}^{8} Presence(attribute_n) \quad (1)$$

$$Presence(attribute_n) = \begin{cases} 1, attribute \notin metadata \\ 2, attribute \in metadata \end{cases}$$

The presence of attributes has a score of 2, while the absence of attributes has a score of 1 (Alomar et al., 2016). The maximum MS value is 16 if all the attributes are extracted from the news article. Similarly, the minimum MS value is 8 when no metadata attributes are extracted. The metadata of the news article are extracted through the *BeautifulSoup* (Richardson, 2007) Python library. It is used for web scrapping by parsing HTML (Hypertext Markup Language) and XML (Extensible Markup Language) content to extract its elements, values and many other attributes.

3.1.2 News content

Online news content contains information on one or more events that occur in the form of paragraphs, media and references of other linked information available electronically to the public (Westerman et al., 2014). The content of the news article is ensured by means of currency, timeliness, relevance, accuracy and its impact (Mandalios, 2013). To analyze the content of the news source, the data quality attributes available in the content are extracted to quantify the data quality score of the content. Extraction of the quality attributes from the content, the content score (CS) calculation of each attribute and the cumulative score are achieved by automated processes. There are different content attributes considered for the assessment of the content of the news source (Alomar et al., 2016). However, it is observed that some of them are mostly unavailable in the news sources. The quality attributes shortlisted to assess content include accessibility, relevance, accuracy, clarity, timeliness and reputation. The CS is computed by the following formulas (Alomar et al., 2016):

$$CS = \sum_{n=1}^{6} Presence(attribute_n) \tag{2}$$

$$Presence(attribute_n) = \begin{cases} 1, Not\ available \\ 2, Partially\ available \\ 3, Available \end{cases}$$

3.1.2.1 Accessibility

The preliminary step of analyzing content is to access online news sources. This ensures that the online news source is available and accessed without any barriers. Moreover, it is also possible that it is available but with restricted access such that it is not possible to access any browser or external tools. In some cases, it is also possible that online news sources are unavailable for future use in digital form.

3.1.2.2 Relevance

Relevance is the most important attribute of the content quality of the online source. In the context of EBS, it is a dependent variable depending upon three further attributes, i.e., affected hosts, an agent that affects the host and the location of the affected host. Furthermore, it is also possible to predict the agent using epidemiological intelligence libraries if it is unavailable. *The Spacy* (Vasiliev, 2020) natural language processing (NLP) Python library is used to perform named entity recognition (NER) to extract locations, hosts and agents. Some examples of the hosts and agents of avian influenza are chickens, pigs, horses, ducks, gooses, etc. and H5N8, H5N1, highly pathogenic avian influenza, etc.

3.1.2.3 Accuracy

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 provides information about any health risk, outbreak information or the number of cases. Alternatively, poor relevance can have poor accuracy but not vice versa. Outbreak information is extracted using *Spacy* (Vasiliev, 2020) by validated outbreak-related keyword tokens in the content. A number of cases are extracted using the pattern-based NLP technique.

3.1.2.4 Clarity

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 (Alomar et al., 2016). Good clarity exists if the subject, description, type, etc. are available in the metadata of the news article.

3.1.2.5 Timeliness

It is important that the content of the news article relates to the current context of the events. Otherwise, the claims may not be considered, or it may be a wrong interpretation. Timeliness is the time of an outbreak saved by detection in EBS relative to the onset of the outbreak (Jafarpour et

al., 2015). Furthermore, timeliness (days) is calculated by the following equation:

$$Timeliness[days] = T_{alarm} - T_{onset} \quad (3)$$

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.

3.1.2.6 Reputation

The reputation of news sources is extracted using the *MediaRank* (Ye and Skiena, 2019) algorithm, which is calculated on multiple factors, i.e., popularity, peer reputation, reporting bias and breadth and bottom-line pressure. For example, the general reputation ranking using *MediaRank* (Ye and Skiena, 2019) of nytimes is '1' and BBC is '5'. Therefore, the general reputation of the news source has an impact on the content quality, as it is computed by considering multiple factors.

3.1.3 The Epidemiological Entity Extraction Approach (E3A)

Event extraction and early warning detection are the key components of EBS (Organization and others, 2014). An event is a verified set of processed epidemiological information of an outbreak (Arsevska et al., 2018). It contains attributes such as location, occurrence date associated with epidemiological entities such as disease or unknown syndrome, symptoms, hosts, agents, etc. (Arsevska et al., 2018). More precisely, this information is available in text in the form of spatiotemporal entities (when, where) and epidemiological entities, i.e., disease, host, agent, symptoms, etc. Furthermore, these attributes are extracted from text using NLP techniques.

In the E3A, 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 from state-of-the-art name-entity recognition (NER) techniques. These named entities are extracted and classified into four categories: spatial, temporal, thematic and other entities. A pattern-based text mining approach is used for extracting and classifying thematic entities such as hosts (e.g., birds, pigs, horses, etc.) that are affected by a disease and variants of different agents (e.g., H5N1, H5N8, HPAI, etc.). After extracting named entities from the title and content of news articles, a weighted-entity approach is proposed for quantifying their epidemiological context in relation to their corresponding title and content. The resulting spatial, temporal and thematic entities that are recognized as relevant entities in the context of a particular EBS attempt (i.e., specific to proposed work) are termed "RelevantEntities (REs)". 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 the title is considered the most important element in text mining approaches used for studying the relevancy. However, a weight of 1.5 is assigned to each RE based on their occurrences in the content of the news article. Moreover, the remaining identified named entities are labeled "OtherEntities (OEs)". Last, a weight of 1 is assigned to each OE regardless of their occurrences in the title and content of the news articles. The E3A score (E3S) is calculated to quantify the context of news articles using the following formula:

$$EntityWeight_n = \begin{cases} 2, RE \in Title\ Sentence \\ 1.5, RE \in Content \\ 1, OE \end{cases}$$

$$E3S = \sum_{i=1}^{n} RE\ Weight/TE\ Weight \quad (4)$$

"RelevantEntityWeight" is the sum of the weight of REs (REs extracted from title and content of the news article). However, "TotalEntityWeight" is the sum of the weight of all entities (REs and OEs) occurring in the title and content of the news article. The E3A is applied to 2 groups of news sources to filter the news articles into relevant and irrelevant categories, which is discussed in Section 4. The occurrences of REs related to the epidemiological context of news sources that are extracted during the experiments are shown using the word cloud in Fig. 2. The word cloud visualization method is chosen from the existing literature because it provides a simple way to communicate the most frequently used relevant words (i.e., REs in our case from the news articles) using different font sizes indicating their occurrence frequency (Lohmann et al., 2015). The identified REs, as shown in Fig. 2, serve as a starting point for further analysis, as stated in Section 4.

In the next subsections, we detail the concept

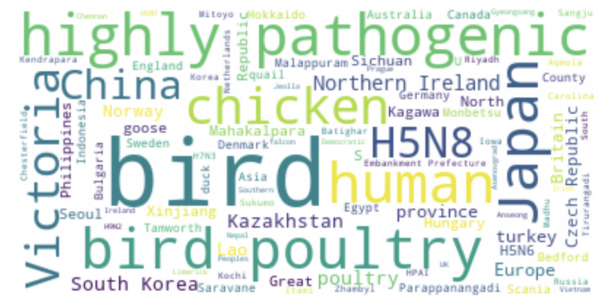

Figure 2: Word Cloud of Extracted REs

of each type of entity as well as the techniques adopted to extract them from news.

3.1.3.1 Spatial Entity

A spatial entity is a geographical or spatial location available in the text (Leidner and Lieberman, 2011). Spatial entities are extracted from the text by using the *Spacy* (Vasiliev, 2020) natural language processing (NLP) Python library, which is mainly used for tokenization, named-entity recognition, etc. Examples of spatial entities include Paris, Rome, Berlin, etc.

3.1.3.2 Temporal Entity

Information on date, time and duration available in the text is known as a temporal entity (Pustejovsky et al., 2003). Temporal entities are important for relating the occurrence of events in news sources. In our proposal, *SUTime* (Chang and Manning, 2012) is used for extracting temporal expressions and temporal entities. Some examples of temporal entities include 'last year', 'this Friday' and 'now', which are resolved into '2020', '2021-05-28' and '2021-05-28', respectively.

3.1.3.3 Thematic Entity

Important information related to our context (in particular to health risks) available in the text is known as thematic entities (Cohen and Hersh, 2005). In our approach, these entities are affected hosts and agents that affect the hosts. Thematic entities in the text are extracted using the *Spacy* (Vasiliev, 2020) Python library with an extended pattern-based approach to extract epidemiological entities. Epidemiological entities are the terms and concepts related to infectious diseases such as hosts, agents, variants and vectors. A pattern-based approach is used to extract named entities (epidemiological entities) in text according to defined patterns. A few examples of thematic entities include birds, H5N1, HPAI, highly pathogenic avian influenza, etc.

3.1.3.4 Other Entity

An entity that never belongs to a spatial entity, temporal entity or thematic entity is called an other entity. A few examples of these entities include quantity: 20, percent: 65%, org: Google, money: 20,000$ and cardinal: three.

4 Evaluation

The purpose of the proposed approach is to filter the relevant news sources by data quality scores. Currently, the proposed work is validated by two groups of news articles, i.e., outbreak-related and irrelevant articles. The ultimate goal of the proposed work is to prioritize outbreak-related news and discard irrelevant news immediately. The details of each group of news items are described in section 4.1.

4.1 Dataset

The dataset of news articles is selected for validation of proposed work having '75' news articles in which the first '25' articles of the dataset contain information about the avian-influenza disease outbreaks, the next '25' articles contain general information (control measures, economic influences, etc.) about avian influenza disease, and the last '25' articles are not health-related news articles. The dataset is small; however, it is more specific to the event outbreak information of avian influenza disease. The preliminary work methodology is validated by a small dataset labeled by experts, whereas subsequent work will be validated with a large dataset labeled by experts. This unstructured information is available in news articles in the form of locations, dates of events with epidemiological entities such as disease or unknown syndrome, symptoms, hosts, and agents. More precisely, this information in news article datasets is used to formulate the spatial-temporal aspects (when, where) and epidemiological aspects, i.e., disease, host, agent, symptoms of the events. Each row in the dataset has a title, text, URL and manually annotated group of news sources. There are two main groups of news articles in the dataset, i.e., irrelevant news (news not related to the health situation) and relevant news (news related to disease outbreaks), which is prioritized by the proposed framework based on quality scores. General information news articles are considered irrelevant in strict evaluation approaches and relevant

Source	DQS	MS	CS	E3S
Outbreak News Today	0.88	0.94	0.89	0.82
Reuters	0.85	0.94	0.89	0.71
The Japan Times	0.85	0.94	0.89	0.71
nippon.com	0.85	0.94	0.89	0.72
The Northern Daily Leader	0.84	0.94	0.89	0.69

Table 1: Top 5 DQS sources in Relevant group

Source	DQS	MS	CS	E3S
CNN	0.72	0.88	0.78	0.51
Aljazeera	0.72	0.81	0.67	0.67
Daily Mail UK	0.88	0.66	0.67	0.63
The Guardian	0.70	0.94	0.67	0.50
CNBC	0.69	0.94	0.61	0.52

Table 2: Top 5 DQS sources in Irrelevant group

in flexible approaches. Furthermore, a news article dataset is available at `http://shorturl.at/auABP`.

4.2 Results

Table 1 shows the top 5 data quality score (DQS) sources of the "relevant" group. *MS* is the metadata score of the news source, *CS* is the content score of the news source, *E3S* is the epidemiological entity extraction score of the news source, and *DQS* is the data quality score computed as the average of the MS, CS and E3S.

Similarly, Table 2 shows the top 5 data quality score (DQS) sources of the "irrelevant" group. The results clearly differentiate the relevant news and irrelevant news by comparing the DQSs.

The results are produced using two evaluation approaches: 1) a strict approach (articles with general information on avian influenza are considered irrelevant) and 2) a flexible approach (articles with general information are considered relevant). The performance evaluation of the proposed work for classifying news articles is based on the classification metrics (Lever et al., 2016). Classification metrics are calculated from true positives (TPs), i.e., relevant news identified as relevant, false positives (FPs), i.e., irrelevant news identified as relevant, false negatives (FNs), i.e., relevant news identified as irrelevant, and true negatives (TNs), i.e., irrelevant news identified as irrelevant.

The results of the news article dataset, which includes quality attribute values, *MS*, *CS* and *E3S*, are available at `http://shorturl.at/foDUV`. For every component score, i.e., MS, CS and E3S, a score of **0.72** and higher is considered a significant quality score for filtering the news articles. A score of 0.72 was considered relevant

Score Type	Precision	Recall	F-Score
MS	0.44	0.96	0.60
CS	0.53	1	0.69
E3S	0.74	0.56	0.64
DQS	0.62	0.96	0.75

Table 3: Strict Approach

Score Type	Precision	Recall	F-Score
MS	0.68	0.80	0.74
CS	0.96	0.90	0.93
E3S	1	0.32	0.48
DQS	0.93	0.76	0.84

Table 4: Flexible Approach

by validation with the manual labels of news articles. The precision, recall and F-score for each DQM using a strict approach are detailed in Table 3. Similarly, precision, recall and F-score for each DQM using a flexible approach are given in Table 4.

The overall performances of different quality measures, i.e., MS, CS, E3S and DQS, with a strict evaluation approach are shown in Figure 3. Similarly, performances with a flexible approach are shown in Figure 4.

In the strict approach, as shown in Table 3, the DQS computed from the average of MS, CS and E3S produced significant improvements in terms of precision, recall and F-Score compared to individual MS, CS and E3S values. Similarly, the same pattern is observed in the flexible approach, as shown in Table 4, with a notable difference in precision, recall and F-score of the DQSs from the individual scores. However, DQSs in both approaches have a higher value than E3S in terms of recall and F-score, except for the precision measure. Conclusively, due to flexibility and dependency on other components, the false positive rate of DQSs is slightly greater than that of E3Ss.

5 Discussion

The quality of available information in online news sources is a mandatory prerequisite for event-based surveillance systems (Edelstein et al., 2018). Thus, one can benefit from using this approach by marking irrelevant news articles retrieved and performing further analysis on relevant news articles. To filter out relevant news articles based on quality attributes, the approach extends the baseline work of Alomar et al. (Alomar et al., 2016) to take into account all contextual information attributes. This work is currently val-

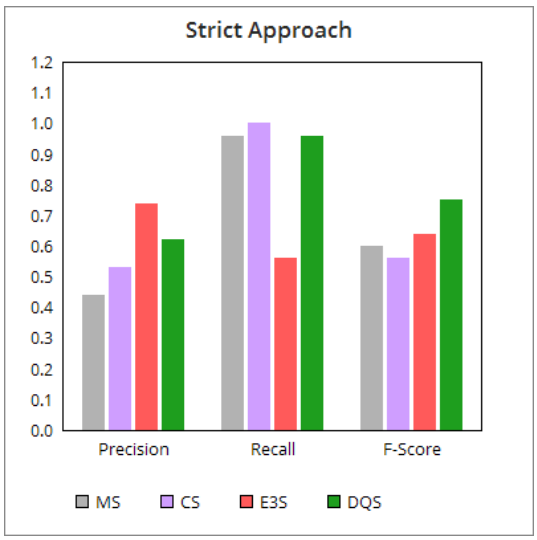

Figure 3: Overall performance - Strict Approach

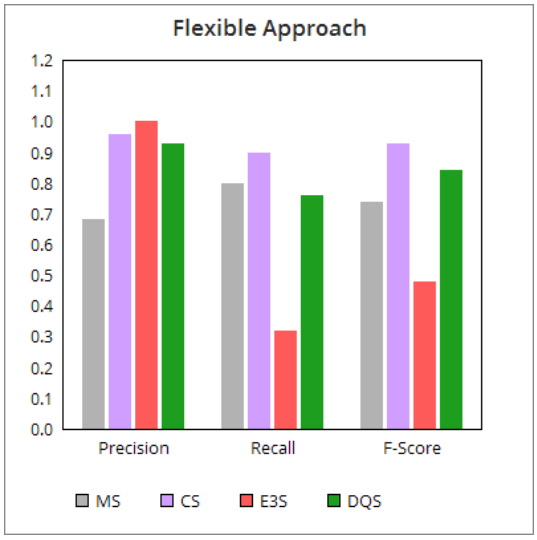

Figure 4: Overall performance - Flexible Approach

idated with a smaller dataset with a disease case study of avian influenza for a proof of concept. One can also validate this approach with a large dataset with avian influenza-related news articles by filtering relevant news articles using the proposed approach.

In the proposed E3A, the weight of the relevant entities is adjusted to determine the context in the news articles. This approach, by weighting the relevant context in news articles, helps filter relevant news in addition to the other components, i.e., metadata and content analysis. Conclusively, the proposed approach produced significant results for the case study of avian influenza. However, it has not been validated with other infectious diseases, but one can try it out with more infectious disease case studies by enhancing the extraction of epidemiological information for those diseases.

6 Conclusion

The proposed research focused on the data quality score computed from the metadata score (MS), content score (CS) with a combination of the epidemiological entity extraction score (E3S) and finally data quality score (DQS), which is the average of the MS, CS and E3S for filtering relevant news articles. The proposed approach is validated using two evaluation protocols, i.e., 1) a strict one and 2) a flexible one. With the strict approach, news is categorized into relevant and irrelevant classes with a precision of 0.62, a recall of 0.96 and an F-score of 0.75. The flexible approach was categorized with a precision of 0.93, a recall of 0.76 and an F-score of 0.84. The MS represents the key attribute aspect enrichment, the CS represents the accuracy, relevance, authority and currency of the content, and the E3S validates the epidemiological context in the news article. The combination of these components resulted in significant improvement in filtering relevant news articles.

Currently, experiments are being conducted on avian-influenza disease news article datasets to filter relevant and irrelevant news. In future work, more data will be integrated in collaboration with experts to extend the avian-influenza dataset and the associated experiments. It could be interesting to investigate and further classify the relevant news articles not only in terms of event outbreaks but also for other specific classes, such as control measures and economic impacts. Further consideration could be possible to analyze non-English, i.e., French, Spanish, etc. news articles and add the capabilities to extract epidemiological information.

Acknowledgments

This study was partially funded by EU grant 874850 MOOD and is catalogued as MOOD023. The contents of this publication are the sole responsibility of the authors and do not necessarily reflect the views of the European Commission.

References

Oscar Alomar, Assumpció Batlle, Josep Maria Brunetti, Roberto García, Rosa Gil, Toni Granollers,

Sara Jiménez, Amparo Laviña, Carme Reverté, Jordi Riudavets, et al. 2016. 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.

Elena Arsevska, Mathieu Roche, Sylvain Falala, Renaud Lancelot, David Chavernac, Pascal Hendrikx, and Barbara Dufour. 2016. Monitoring disease outbreak events on the web using text-mining approach and domain expert knowledge. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3407–3411.

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

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

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

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

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

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

Herman Anthony Carneiro and Eleftherios Mylonakis. 2009. Google trends: a web-based tool for real-time surveillance of disease outbreaks. *Clinical infectious diseases*, 49(10):1557–1564.

Kenrick D Cato, Bevin Cohen, and Elaine Larson. 2015. 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.

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

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

Aaron M Cohen and William R Hersh. 2005. A survey of current work in biomedical text mining. *Briefings in bioinformatics*, 6(1):57–71.

Michael Edelstein, Lisa M Lee, Asha Herten-Crabb, David L Heymann, and David R Harper. 2018. Strengthening global public health surveillance through data and benefit sharing. *Emerging Infectious Diseases*, 24(7):1324.

Mohamed K Elhadad, Kin Fun Li, and Fayez Gebali. 2019. 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.

Marwa Essam and Tamer Elsayed. 2020. 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.

Iris Ganser. 2020. *Evaluation of event-based internet biosurveillance for multi-regional detection of seasonal influenza onset*. Ph.D. thesis, McGill University (Canada).

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

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

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

Mira Kim, Kyunghee Chae, Seungwoo Lee, Hong-Jun Jang, and Sukil Kim. 2020. 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.

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

Jake Lever, Martin Krzywinski, and Naomi Altman. 2016. Classification evaluation (vol 13, pg 603, 2016). *NATURE METHODS*, 13(10):890–890.

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. 2010. Quality of traditional surveillance for public reporting of nosocomial bloodstream infection rates. *JAMA*, 304(18):2035–2041.

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

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

Yoshiko Nozato. 2002. 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*.

World Health Organization et al. 2008. A guide to establishing event-based surveillance. *World Health Organization*.

World Health Organization et al. 2014. 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.

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

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

Leonard Richardson. 2007. Beautiful soup documentation. *Dosegljivo: https://www. crummy. com/software/BeautifulSoup/bs4/doc/.[Dostopano: 7. 7. 2018]*.

Charlotte Rudnik, Thibault Ehrhart, Olivier Ferret, Denis Teyssou, Raphael Troncy, and Xavier Tannier. 2019. 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. Association for Computing Machinery.

Sarah Valentin. 2020. *Extraction and combination of epidemiological information from informal sources for animal infectious diseases surveillance*. Ph.D. thesis, Université Montpellier.

Yuli Vasiliev. 2020. *Natural Language Processing with Python and SpaCy: A Practical Introduction*. No Starch Press.

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

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

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

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

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

Xiaolan Zhu and Susan Gauch. 2000. 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.