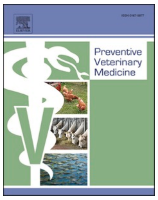
Contents lists available at ScienceDirect

Preventive Veterinary Medicine

journal homepage: www.elsevier.com/locate/prevetmed



Animal disease surveillance: How to represent textual data for classifying epidemiological information

Sarah Valentin [a,b,c,d], Rémy Decoupes [c], Renaud Lancelot [a,b], Mathieu Roche [a,c,*,1]

[a] *CIRAD, F-34398 Montpellier, France*
[b] *ASTRE, Univ Montpellier, CIRAD, INRAE, Montpellier, France*
[c] *TETIS, Univ Montpellier, AgroParisTech, CIRAD, CNRS, INRAE, Montpellier, France*
[d] *Département de Biologie, Université de Sherbrooke, Sherbrooke, Québec, Canada*

ARTICLE INFO

Keywords:
Animal disease surveillance
Classification
Word embedding

ABSTRACT

The value of informal sources in increasing the timeliness of disease outbreak detection and providing detailed epidemiological information in the early warning and preparedness context is recognized. This study evaluates machine learning methods for classifying information from animal disease-related news at a fine-grained level (i.e., epidemiological topic). We compare two textual representations, the bag-of-words method and a distributional approach, i.e., word embeddings. Both representations performed well for binary relevance classification (F-measure of 0.839 and 0.871, respectively). Bag-of-words representation was outperformed by word embedding representation for classifying sentences into fine-grained epidemiological topics (F-measure of 0.745). Our results suggest that the word embedding approach is of interest in the context of low-frequency classes in a specialized domain. However, this representation did not bring significant performance improvements for binary relevance classification, indicating that the textual representation should be adapted to each classification task.

1. Introduction

The ability to rapidly recognize emerging and re-emerging animal infectious diseases is a critical global health priority. Early warning is crucial for the quick implementation of effective control strategies at global and local levels (Heymann and Rodier, 2001). In recent decades, several outbreaks have highlighted the limitations of conventional disease surveillance, which is hampered by delayed detection and latency of the communication channels (Ben Jebara and Shimshony, 2006). The growing availability of digital data represents an unprecedented source of real-time disease information for epidemic intelligence (EI) (Paquet et al., 2006). Online news, social media and electronic health records are among the so-called informal sources that have proven to be valuable sources of disease information (Soto et al., 2008; Wilson and Brownstein, 2009; Dion et al., 2015; Bahk et al., 2015). Their mainstreaming into surveillance systems via the concept of event-based surveillance has been a game-changer for disease surveillance and control. While the earliest applications have focused on human health, event-based surveillance has also been successfully applied to both animal and zoonotic diseases (Arsevska et al., 2018).

Informal information sources are diverse in their spectrum (e.g., online news and social media messages), but they all share information in textual format. Peculiarities of textual data include linguistic ambiguities, redundant and noisy information, and a lack of normalization. In addition, daily amounts of such information can rapidly overwhelm surveillance systems, including a step of moderation by experts. Event-based surveillance (EBS) systems, such as HealthMap or GPHIN, have thus been developed to collect and process the continuous flow of informal information. Such systems increasingly marshal text-mining methods to alleviate the amount of manually curated free text (Hartley et al., 2010). Text mining, which combines natural language processing (NLP) and data mining techniques, enables free text conversion into a computer-readable format (Hearst, 1999). The pipelines of EBS systems include four steps, from data collection to the extraction of relevant pieces of epidemiological information (Fig. 1). Each step can be performed manually, semiautomatically or automatically according to the system.

In the animal health domain, the final extraction of epidemiological information (or entities) aims at identifying, among others, the disease names, species, locations, and dates related to a specific outbreak.

[*] Corresponding author at: CIRAD, F-34398 Montpellier, France.
E-mail address: mathieu.roche@cirad.fr (M. Roche).

[1] Postal address: CIRAD, Campus de Lavalette, Avenue Agropolis, 34398 Montpellier Cedex 5, France.

https://doi.org/10.1016/j.prevetmed.2023.105932
Received 1 April 2022; Received in revised form 7 April 2023; Accepted 10 May 2023
Available online 12 May 2023

Before this step, a classification step is usually performed to filter the relevant documents (e.g., online news). Classification is a crucial step of animal disease surveillance systems, as it avoids overwhelming the systems with irrelevant data (e.g., a review about a disease). The classification approaches integrated into the EBS systems differ regarding the number of categories used to label the documents, the type of classification method, and the kind of moderation. Several tools, such as HealthMap, GPHIN and PADI-web (Brownstein et al., 2008; Carter et al., 2020; Valentin et al., 2020b), rely on classifiers trained on manually labeled datasets to automatically learn rules to label the retrieved news articles (so-called supervised classification). The classifiers include, for instance, naïve Bayes, the support vector machine (SVM), and more recently, deep-learning classifiers such as bidirectional long short-term memory recurrent neural networks (Kim et al., 2020).

EBS system classification frames usually assign one category per document (i.e., relevant/irrelevant). However, there is a challenging heterogeneity of topics and relevance scores across sentences. For instance, news articles that report an outbreak often also describe outbreak control measures or economic impacts, share information about the outbreak source or draw attention to a given area at risk. Those elements may be of relevance to EI teams to assess risks associated with the occurrence of an outbreak. Conversely, some sections do not contain any useful information, which can result in noisy entity extraction.

In this context, we stress the need for an intermediate step between document-based classification and the extraction of epidemiological entities. This step consists of i) removing irrelevant content (i.e., sentences) from relevant documents and ii) identifying relevant epidemiological topics, which we refer to as fine-grained information.

Several applications can be derived from this approach in EBS pipelines. First, sentence-based classification may enhance the performance of event-extraction tasks by identifying event-related sentences (Gella and DuongThanh, 2012; Naughton et al., 2010). Performing event extraction on a subset of relevant sentences would decrease the risk of extracting epidemiological entities (e.g., dates and locations) not related to an event. In addition, sentences related to transmission pathways can be manually or automatically compared to current disease knowledge to identify the emergence of a new transmission pathway. Sentence-based classification is an alternative approach to increase the performance of document-based classification. Especially in the context where event-related information appears within a few sentences, such a strategy was recently applied to the classification of infectious disease occurrences from online textual sources (Kim et al., 2020). The authors proposed an approach of classifying each sentence first and then merging the results of each sentence classification to classify the document. They obtained better performance through the sentence-based approach compared to a document-unit learning classifier.

Textual classification involves using statistical learning models to classify text (e.g., a whole document and a sentence) into specific sets of categories. The classification is supervised when models are trained on instances whose labels are known (i.e., annotated by domain experts) (Witten et al., 2016). Models (i.e., classifiers) are fitted on annotated instances during the training step, which includes two steps: textual vectorization and classification. Textual vectorization (hereafter referred to as "representation") converts textual data into a machine-readable format. In the classical bag-of-words (BOW) representation, each document is transformed into a sparse vector where each dimension is an explicit feature, i.e., a word. In recent years, representations based on deep neural networks (word embeddings) have been introduced as an alternative to bag-of-words to better capture syntax and semantic information from text. Popular models include Word2Vec (Mikolov et al., 2013a), GloVe (Pennington et al., 2014), and BERT (Devlin et al., 2019b). In such approaches, documents are in an implicit space where they are represented as a dense vector (Mikolov et al., 2013a). Word embedding models have yet to be applied to several linguistic tasks in the disease surveillance domain, including disease taxonomy development (Ghosh et al., 2016), epidemiological feature extraction from WHO reports (Ghosh et al., 2017) and veterinary necropsy report classification (Bollig et al., 2020).

Compared with document classification, the classification of short text (such as sentences) is more challenging because of the lack of contextual information (Song et al., 2014; Chen et al., 2019). Such shortcomings may hamper the performance of classic textual representations that rely on the use of word presence and/or frequency (e.g., bag-of-words), suggesting the need for models able to better capture the word semantics, such as word embeddings. In this paper, we propose integrating the Word2Vec model in the context of a specialized domain, i.e., text-based animal disease surveillance. More precisely, we aim to evaluate what textual representation to implement for classifying sentences from online news to identify epidemiological information, including the BOW and word embedding approaches.

2. Materials and methods

2.1. Corpus and classification task

We used a publicly available corpus of news articles that was annotated by four epidemiologists following specific guidelines (Valentin et al., 2019). In this annotation framework, news articles are split into sentences, and each sentence has two levels of annotation (i.e., two labels) (Valentin et al., 2021). The first level aims at identifying if the sentence contains any relevant epidemiological information, containing five levels (current, hypothetical or past events, general and irrelevant sentences). In this study, we aggregated sentences labeled irrelevant (i.e., disease-unrelated general facts) and general (i.e., general information about a disease or a pathogen) to create a group of irrelevant sentences.

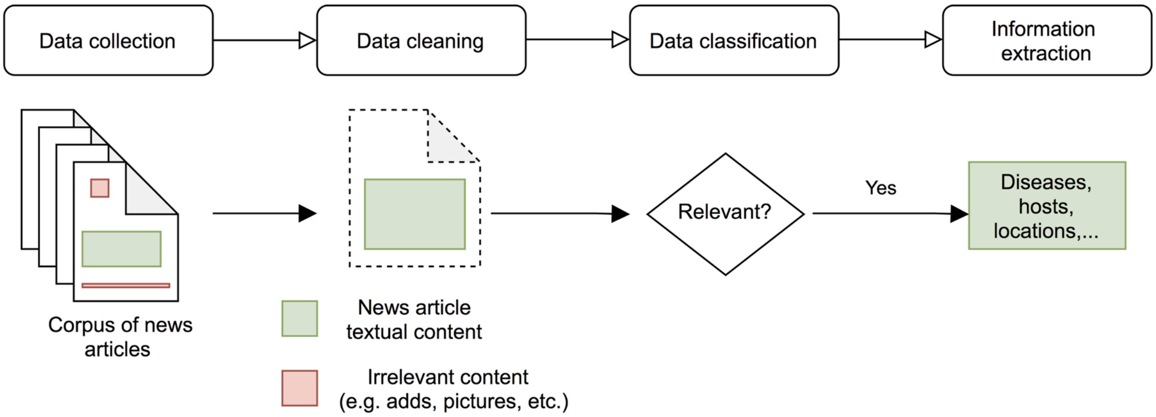

Data collection
Data cleaning
Data classification
Information extraction
Corpus of news articles
Relevant?
Yes
Diseases, hosts, locations,...
News article textual content
Irrelevant content (e.g. adds, pictures, etc.)

Fig. 1. Pipeline of news article processing in event-based surveillance systems.

The other labels (current, hypothetical and past events) are aggregated into the relevant category. The second annotation level characterizes the relevant sentences' epidemiological topic (fine-grained information), as shown in Table 1.

In this study, we adopted the supervised classification paradigm using the previously described corpus of sentences as the training dataset. The two levels form two consecutive classification tasks: i) classification of the relevance status and ii) topic classification of the relevant sentences (Fig. 2). The first classification is binary (sentences are either relevant or irrelevant), while the sentence is multiclass (i.e., sentences belong to one of the six topic categories). To evaluate retrieval methods on sufficient class sizes, we increased the initial annotated corpus (32 news articles, 486 sentences - 10,247 words) with 758 sentences (16,417 words). We obtained a final corpus containing 1244 sentences, among which 296 sentences were irrelevant. Hence, the subset of sentences for topic classification consisted of 948 sentences (21,753 words).

Even if still modest in size, our corpus is specific regarding both its domain (i.e., animal health) and its nature (i.e., online news articles). Therefore, this corpus type is more specific than the benchmark corpus traditionally used in state-of-the-art approaches in the biomedical NLP domain (Huang and Lu, 2015).

Our objective is to fit a classification model able to correctly identify both the relevance status and the topic of unlabeled sentences.

2.2. *Textual representation*

The transformation of textual data into a vector-space representation assumes that a document from a corpus can be represented as a numeric vector derived from the terms it contains (Salton, 1971). *In fine*, the closeness of two document vectors in the vector space should reflect their semantic similarity. In this work, we compare two types of vector-space representations, i.e., the traditional bag-of-words and the word embedding representations. In the following subsections, we outline the construction of each textual representation. For both types of representations, we evaluated three types of textual preprocessing steps: no preprocessing (P1), lowercase (P2), and lowercase plus lemmatization (P3), which consists of converting the words into their inflected forms (e.g., plural to singular form).

2.2.1. *Bag-of-words representation*

In the BOW representation, each document d (here, each sentence) is represented by an n-dimensional vector where each component w_{id} represents the absence or presence of a feature (term) i in the document and n is the length of the vocabulary (Zhang et al., 2010). If the feature i occurs in the document, the feature weight w_{id} has a nonzero value. As a weight, we used the term frequency-inverse document frequency ($TF - IDF$), as shown in Eq. (1). $TF - IDF$ is the product of term frequency TF_{id} - the frequency of the word in the document - and the inverse document frequency IDF_i, given by Eqs. (2) and (3):

$$w_{id} = TF_{id} \times IDF_i \tag{1}$$

$$TF_{id} = \frac{n_{id}}{n_d} \tag{2}$$

$$IDF_i = log\left(\frac{N}{D_i}\right) \tag{3}$$

where .

- n_{id} is the number of times that the term i appears in document d,
- n_d is the total number of terms in d,
- N is the total number of documents in the corpus, and
- D_i is the number of documents that contain i.

Terms with the highest $TF - IDF$ values are distinctively more frequent in a document compared to the collection of documents (Salton and Buckley, 1988).

Table 1
Classification tasks.

Class	Number of sentences	Example
Descriptive epidemiology (DE)	401	Cases of African swine fever (ASF) have been recorded in Odesa and Mykolaiv regions.
Protection and control measures (PCM)	293	All the infected animals have been killed, and the area has been disinfected.
Concern and risk factors (CRF)	105	Additional outbreaks of African swine fever are likely to occur in China.
Transmission pathway (TP)	69	The authorities suggest that the highly contagious virus might have been spread by a river.
Economic and political consequences (EPC)	53	Financial losses due to ASF could amount to 17 million to Latvia's industry in 2017.
Distribution (DI)	27	27 94 poultry farms in Taiwan have been infected by avian flu so far this year.
Total	948	

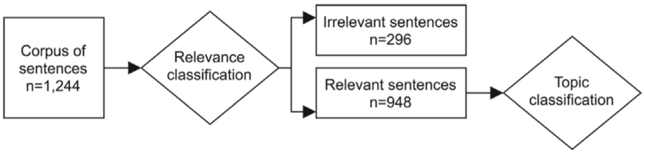


Fig. 2. Classification tasks.

Bag-of-words is an easy-to-understand and effective representation used to convert textual documents into vectors. However, it has several limitations, including its sparsity (Brownlee, 2017; Zhao and Mao, 2018). The BOW representation leads to highly sparse vectors (i.e., most of the vector elements have zero value since a document only contains a very small portion of all of the vocabulary). This may result in computational complexity while drowning out information. Moreover, the BOW representation overlooks the grammar and word order in a document, as reflected by the "bag" concept. The context of the terms is discarded even though it provides meaningful information regarding the semantics of terms, such as synonymy. For instance, the BOW representation may not effectively capture the closeness of semantically similar documents with different term usages, as they are converted to very different vectors.

2.2.2. *Word embedding representation*

Word embedding methods produce word representations corresponding to dense real-valued vectors in a vector space (Torregrossa et al., 2021). Vector values are learned according to the context in which the word appears, based on the assumption that words that frequently appear in the same context (i.e., surrounded by the same words) tend to have the same meaning (Goldberg, 2017). In most approaches, the context corresponds to the window of neighboring words, which is a configurable parameter. For instance, in health-related news, the verbs "declared" and "reported" are typically used in the same types of sentences (e.g., "France declared/reported an outbreak of foot-and-mouth disease"). While the traditional bag-of-words representation will encode the verbs "declared" and "reported" as two distinct features, a word embedding representation may capture their semantic closeness.

Contrary to BOW, word embedding representations are obtained by a learning process on a large corpus. This learning process is distinct from subsequent NLP tasks for which embeddings are used, namely, text classification in our study. Several pretrained word embeddings are publicly available, but training a word embedding model on text specific to the target domain has been shown to improve performance (Pyysalo et al., 2015). In this study, we thus decided to compare a pretrained model (w2v-G) with different custom-trained models (w2v-P1, w2v-P2, and w2v-P3), whose characteristics are summarized in Table 2.

Two types of parameters can influence the quality of vectors when training a word embedding model: (i) the preprocessing steps of the training dataset and (ii) the model parameters. As prepreocessing steps, we evaluated the influence of lowercasing and lemmatization. In this study, we focused only on the word2vec model. The word2vec model consists of a 2-layer neural network and was proposed with two architectural variants: the continuous bag-of-words (CBOW) and the continuous skip-gram. The CBOW architecture predicts a target word based on its context, while the skip-gram model attempts to predict a target context using a word (Mikolov et al., 2013b). As each method has its own advantages, we compared the performance of each architecture. We evaluated two dimensions for the trained vectors: 100,000, which is the default length implemented in the gensim library, and 300,000, which is the most commonly used dimensionality in various studies (Mikolov et al., 2017, 2013a; Pennington et al., 2014). For simplicity, both dimensions are further referred to as "100" and "300". We used the default parameter for the window size (5 words).

We trained the custom models on a dataset of news articles dealing with the animal health domain. This dataset is called Epi-Animal. The training set length was 33,417,501 words.

The word2vec pretrained model was trained on a 3 billion-word corpus from the Google News corpus[2] with the CBOW architecture.

Each sentence vector was computed by calculating the average of the sentence's word embeddings. As proposed by (De Boom et al., 2016; Krzeszewska et al., 2022), each word embedding was leveraged by the $TF-IDF$ word value. Finally, each sentence is represented by a vector that pools the information of all of the words.

2.3. Classifiers

We compared two classifiers that are widely used for text classification, i.e., support vector machines and the multilayer perceptron.

1. A support vector machine (SVM) is a nonprobabilistic and linear classification technique. SVMs have been widely used for text classification, including small text such as sentences (Khoo et al., 2006; Zhang and Liu, 2007) and tweets (Go et al., 2009). A SVM can perform well regarding important textual data vector properties, which contain few irrelevant features (Joachims, 1998). We used a linear kernel parameter (linear SVM) classifier, as linear kernels can perform well with textual data (Uysal and Gunal, 2014; Kumar et al., 2010).
2. The multilayer perceptron (MLP) is an artificial neural network (ANN)-type classifier. ANN classifiers were shown to perform well when combined with word embedding representations (Agibetov et al., 2018; Mandelbaum and Shalev, 2016). We used the default parameters implemented in the scikit-learn library (Pedregosa et al., 2011), i.e., 1 hidden layer, 100 hidden units and a rectified linear unit (ReLU) activation function.
3. As this study focuses on the evaluation of the textual representation, we did not fine-tune the classifiers. We are nevertheless aware that fine-tuning can improve the performances of each classifier, and our result must be interpreted in the context of the set of parameters cited above.

Preprocessing, word embedding model training, classification and the evaluation pipeline were performed using the scikit-learn, NLTK and gensim libraries (Python) (Pedregosa et al., 2011; Bird and Loper, 2004; Rehur and Sojka, 2010). The code is freely available at: https://github.com/SarahVal/EpiNews-Representation/.

2.4. Evaluation

We estimated the performances of the trained models via the widely used cross-validation method. We used a fold number of 5, as this value was empirically shown to yield test error rate estimates with low variance, while not being hampered by excessively high bias (Hastie et al., 2009).

At each fold, we computed the traditional metrics used in supervised classification, i.e., precision, recall, accuracy, and F-measure. Precision for a given class A corresponds to the proportion of correct sentences classified in class A (Eq. (4)), and recall corresponds to the proportion of sentences belonging to class A that are correctly identified (Eq. (5)):

$$Precision(A) = \frac{number \quad of \quad sentences \quad correctly \quad attributed \quad to \quad class \quad A}{number \quad of \quad sentences \quad attributed \quad to \quad class \quad A} \tag{4}$$

$$Recall(A) = \frac{number \quad of \quad sentences \quad correctly \quad attributed \quad to \quad class \quad A}{total \quad number \quad of \quad sentences \quad belonging \quad to \quad class \quad A} \tag{5}$$

F-measure is the harmonic mean of precision and recall (Eq. (6)).

$$F-measure(A) = \frac{2 \times Precision(A) \times Recall(A)}{Precision(A) + Recall(A)} \tag{6}$$

To calculate the performances over all classes to account for class imbalance, we computed the weighted precision, recall, and F-measure (averaging the support-weighted mean per label). The accuracy corresponds to the proportion of correct predictions over the total predictions.

We conducted two experiments. The first aimed at identifying the best preprocessing steps and parameters for training the word embedding model. The second compared the results of the classifiers and representations detailed above. To evaluate and compare the textual representations obtained, we adopted a two-step evaluation: .

- Selection of the best parameters of the word embedding model for the topic classification of relevant sentences based on the best overall accuracy.
- Comparison of the selected models with the bag-of-words representation for the relevance and topic classification tasks.

3. Results

3.1. Word embedding parameters

We compared the performances of the word embedding-based representations for relevance and topic classification in terms of the weighted F-measure, with both MLP and SVM classifiers (Table 3).

The topic classification yielded lower results than the relevance classification, with a weighted F-measure ranging from 0.666 to 0.745, compared to 0.784–0.871 for relevance classification.

For relevance classification, the w2v-P1 representation obtained from skip-gram architecture and used to feed the MLP classifier achieved a higher F-measure than other parameter combinations, with 100-dimension vectors (0.871). For topic classification, the best models were the w2v-P2 model obtained from the skip-gram architecture used to provide the SVM classifier and the w2v-P1 model trained using the skip-gram architecture used to provide the SVM, with the 100-dimension and 300-dimension vectors, respectively.

Even though the best performances were reached with skip-gram, both algorithms achieved comparatively equal performances among the parameter combinations. Similarly, our results do not reveal any superiority of a vector length among the others. In contrast, the F-measure significantly dropped when preprocessing method P3

[2] https://github.com/mmihaltz/word2vec-GoogleNews-vectors

Table 2
Characteristics of the word embedding models evaluated, including the pre-trained word2vec embeddings (w2v-G) and the models trained on Epi-Animal corpus (w2v-P1, w2v-P2 and w2v-P3).

Representation name	Training parameters		Training corpus			Model output
	Architecture	Pre-trained	Source	Size (tokens)	Pre-processing method	Vocabulary size
w2v-G	CBOW	yes	Google news (generalist)	3 bilions	Figures removed	3,000,000
w2v-P1	CBOW, Skip-gram	no	Epi-animal corpus (specialized)	33 millions	P1: None	378,609
w2v-P2					P2: Lowercase	315,306
w2v-P3					P3: Lowercase and lemmatisation	282,611

Table 3
Performances of relevance and topic classification depending on the pre-processing method (w2v-P1: none, w2v-P2: lowercase, w2v-P3: lowercase and lemmatisation), the word2vec model architecture, the embeddings vector length, and the classifier, in terms of weighted F-measure.

			Relevance classification		Topic classification	
Representation name	Architecture	Classifier	Vector length			
			100	300	100	300
w2v-P1	CBOW	SVM	0.825	0.833	0.729	0.742
		MLP	0.870	0.858	0.722	0.722
	Skip-gram	SVM	0.843	0.841	0.733	**0.745**
		MLP	**0.871**	0.846	0.727	0.731
w2v-P2	CBOW	SVM	0.832	0.842	0.728	0.744
		MLP	0.858	0.864	0.719	0.728
	Skip-gram	SVM	0.828	0.836	**0.745**	0.736
		MLP	0.869	0.845	0.740	0.739
w2v-P3	CBOW	SVM	0.799	0.807	0.691	0.700
		MLP	0.839	0.835	0.674	0.672
	Skip-gram	SVM	0.784	0.796	0.687	0.686
		MLP	0.840	0.809	0.691	0.666

(lowercase and lemmatization) was applied for both classification tasks. The MLP classifier yielded better performance for relevance classification, regardless of the preprocessing method and model architecture, while the performances of the SVM and MLP were less different for topic classification. Nevertheless, except for one combination, the SVM performed better than the MLP for this task.

3.2. Classifiers and representations

In this second experiment, we compared the performances of two textual representations (i.e., bag-of-words or word embedding) for both relevance and topic classification, depending on the pre-processing method and the classifier. As word embedding representations, we selected the 100-length vectors obtained through the skip-gram architecture because they reached the highest weighted F-measure in the previous sections. We compared their performances with those of the pretrained word embeddings (G-Emb), as shown in Fig. 3. Note that the model w2v-G (pre-trained on a general corpus) is shown in the P1 section as none lowercasing nor lemmatization were used for its training.

For relevance classification, bag-of-words and word embedding representations obtained comparable results, except for the combination w2v-MLP, which reached the highest weighted F-measure regardless of the preprocessing method. The performances obtained by the pretrained word embeddings (w2v-G) were comparable to those of the custom-trained embedding. We noted a slight decrease in performance when using the P3 preprocessing method for both BOW and embedding representations.

The results obtained for topic classification exhibited greater differences, with a clear improvement when using custom-trained embeddings, regardless of the preprocessing method. The MLP and SVM classifiers obtained comparable performances with embeddings, while the use of the MLP decreased the performances combined with the BOW

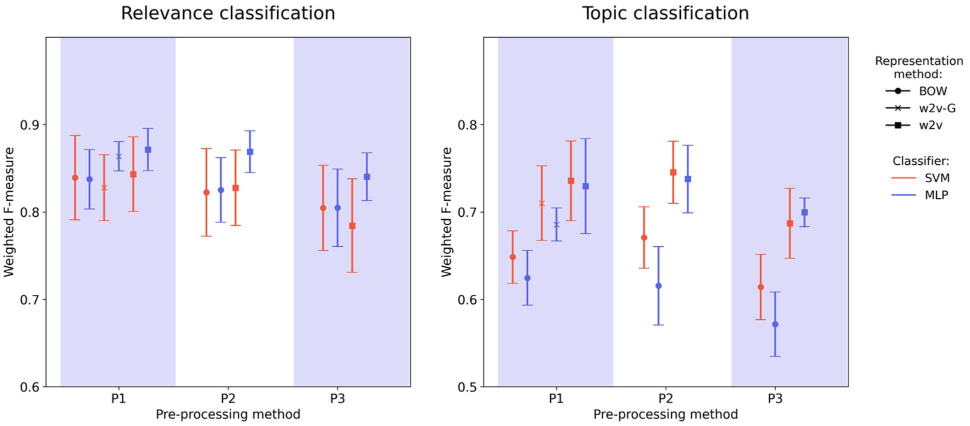


Fig. 3. Performances of relevance and topic classification results based on bag-of-words (BOW) and word embedding representations in terms of weighted F-measure (w2v-G: pretrained word2vec embeddings on a generalist corpus, w2v: word2vec embeddings trained on a specialized corpus). Confidence intervals represent the standard deviation of the F-measure across the five cross-validation folds.

representations. Similar to relevance classification, the P3 representation obtained the lowest F-measure.

Table 4 summarizes the performances of the best BOW and word embedding representations (in terms of weighted F-measure) for each classification task. For relevance classification, the best word embedding representation (w2v-P1-skip-100) brings a gain in the weighted F-measure of 0.005 (SVM classifier) and 0.035 (MLP classifier) compared to the best BOW representation. For topic classification, the best word embedding representation (w2v-P2-skip-100) brings a gain in the weighted F-measure of 0.074 (SVM classifier) and 0.118 (MLP classifier) compared to the best BOW representation.

We analyzed the intraclass classification results and confusion matrices for topic classification, comparing the best representations identified in the previous step (BOW-P2, classified by an SVM classifier, and w2v-P2-skip-100, classified by an MLP classifier (Table 5)). The word embedding outperformed the BOW classification in terms of weighted precision, recall and F-measure in all classes, including underrepresented classes (e.g., concern and risk factors and economic and political consequences). Notably, the classification results were highly heterogeneous between the classes. The F-measures ranged from 0.419 (“distribution” class) to 0.796 (“descriptive epidemiology” class) with the bag-of-words representation. They ranged from 0.500 to 0.841 with the word embedding representation (same classes). With both textual representations, the lowest recall and precision tended to be correlated with the class having the lowest number of instances, such as “distribution” (n = 27) and “economic and political consequences” (n = 53). For this latter class, the word embedding particularly increased the classification performance (F-measure increased by 0.162).

The classification errors were largely due to false classifications into the two majority classes (descriptive epidemiology and protective and control measures) (see Fig. 4). The word embedding representation decreased the number of instances falsely attributed to both classes, with the notable exception of distribution sentences attributed to the descriptive epidemiology class.

4. Discussion

The classification results based on both BOW and word embedding representations showed that the supervised approach performed well in identifying relevant sentences in animal-health-related news (with the highest weighted F-measure values reaching 0.839 and 0.871, respectively). While trained on approximately the same dataset lengths, the performance of document-based classification reported by the EBS system is higher (F-measure of 0.93 for GPHIN (Conway et al., 2009), accuracy of 0.92 for PADI-web on an external dataset (Valentin et al., 2020a)). Sentence classification is more challenging due to sentence sparsity. The classification of epidemiological topics achieved poorer global performances, with the underrepresented classes obtaining the lowest F-measures. This can be explained by the multiclass classification task, which may require richer semantics. Moreover, the topic classification includes underrepresented and low-frequency classes, which trigger classification errors in a supervised framework. This can be a major limitation in practice when retrieving underrepresented classes, such as a transmission pathway or concern and risk factors.

Table 5
Result of topic classification with bag-of-words (BOW-P1) and word embedding (w2v-P2-skip-100) representations, in terms of recall, precision and F-measure. The best performances are shown in bold for each level.

	Bag-of-words			Word embeddings		
Class	Precision	Recall	F-measure	Precision	Recall	F-measure
DE	0.796	0.796	0.796	0.841	0.841	0.841
PCM	0.688	0.669	0.678	0.791	0.737	0.763
CRF	0.449	0.419	0.433	0.646	0.610	0.627
TP	0.514	0.536	0.525	0.536	0.536	0.536
EPC	0.456	0.491	0.473	0.580	0.755	0.656
DI	0.371	0.481	0.419	0.388	0.704	0.500

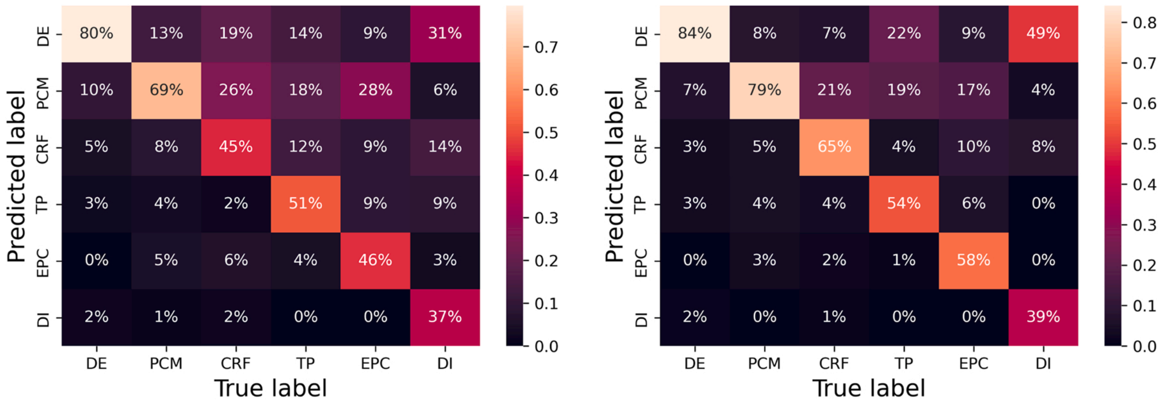
DE: Descriptive epidemiology, PCM: Protection and control measures, CRF: Concern and risk factors, TP: Transmission pathway, EPC: Economic and political consequences, DI: Distribution

Our results weakly suggest that training on word embeddings may overcome this limitation by learning more effectively the semantic meaning of short texts. Embedding models are trained on external datasets, which is contrary to traditional bag-of-words representations. They thus allow the classifiers to generalize more effectively beyond their limited number of training examples (Thapen et al., 2016). Such a feature of word embedding models is of utmost interest when implementing supervised approaches in a specialized domain, such as animal disease surveillance. The model trained on the Epi-Animal corpus achieved slightly better results than the pretrained word2vec model, suggesting that a specialized corpus had a greater value than training on a larger generalist corpus. Further evaluation is needed to validate this result.

For relevance classification, the use of word embeddings did not strongly outperform bag-of-words representations. Previous studies have suggested that word embeddings with traditional classifiers and a sufficient learning dataset size may not have additional value for classification tasks (d’Amato et al., 2017; Bollig et al., 2020). In Bollig et al. (2020), the GloVe embedding model did not outperform the TF-IDF vector model in classifying necropsy reports. Compared to our study, the model was trained on a smaller dataset (1000 reports, forming a 50, 000-word vocabulary). Although these approaches cannot be directly compared since the word embedding models were different, we may hypothesize that the added value of word embeddings also depends on the sparsity of the data classified, i.e., short text such as sentences or longer documents such as reports or news articles. In our results, none of the word2vec training parameters (i.e., architecture and vector length) significantly outperformed the other parameters. Skip-gram is known to be more efficient, with infrequent words and small training datasets (Naili et al., 2017). Hence, we expected this model to perform better than CBOW. The choice of embedding dimension is still an open issue in the literature. It relies on a trade-off between small dimensionality, which may not capture all possible word relations, and large dimensionality, which suffers from overfitting (Yin and Shen, 2018). Our results suggest that 100-length vectors can be used without impacting the overall accuracy of the classification, although this may not be generally

Table 4
Performances of relevance and topic classification based on bag-of-words and word embedding representations, in terms of weighted precision, recall, F-measure and accuracy.

Classification task	Classifier	Textual representation	Precision	Recall	F-measure	Accuracy
Relevance	SVM	BOW-P1	0.838	0.841	0.839	0.841
		w2v-P1-skip-100	0.858	0.837	0.843	0.837
	MLP	BOW-P1	0.837	0.846	0.836	0.846
		w2v-P1-skip-100	**0.870**	**0.873**	**0.871**	**0.873**
Topic	SVM	BOW-P2	0.672	0.670	0.671	0.670
		w2v-P2-skip-100	**0.754**	0.741	**0.745**	0.741
	MLP	BOW-P2	0.636	0.642	0.622	0.642
		w2v-P2-skip-100	0.738	**0.745**	0.740	**0.745**



(a) Bag-of-words.

(b) Word embeddings.

Fig. 4. Confusion matrices for topic classification with the best bag-of-words (BOW-P1) and word embedding (w2v-P2-skip-100) representations. DE: Descriptive epidemiology, PCM: Protection and control measure, CRF: Concern and risk factors, TP: Transmission pathway, EPC: Economic and political consequences, DI: Distribution.

true if word embeddings are used for a different dataset or task. This simpler model could be preferred in terms of its reduced computing time and complexity (for both textual preprocessing and model training steps), which are constraints that can hinder the model applicability of large embedding vectors (Wu et al., 2016). Conversely, the choice of textual preprocessing steps significantly impacted our results, with the use of lemmatization decreasing the performances, highlighting the need to evaluate different preprocessing combinations in classification pipelines. Eventually, our results did not reveal any better performance of one classifier compared to the other when using the model's default values of the hyperparameters. Nonetheless, the classification performances of each classifier can be further optimized through hyperparameter tuning (Elgeldawi et al., 2021).

New word embedding architectures have been recently proposed, such as the BERT model (bidirectional encoder representations from transformers). This model achieved new state-of-the-art results on several NLP tasks, including sentence classification (Devlin et al., 2019a; Torregrossa et al., 2021). BERT produces word representations that are dynamically informed by the words around them (also referred to as "contextualized word embedding"). These word embedding architectures have also been applied to the extraction of fine-grained events from online news (Piskorski et al., 2020), outperforming the BOW representation. We conducted preliminary analysis on our classification task with a BERT-like pretrained model, RoBERTa (Liu et al., 2019). The results indicate a clear improvement in performance. For instance, topic classification with the RoBERTa model reached a weighted accuracy and F-measure of 0.84. In our future work, we plan to combine the best representations related to the dedicated tasks highlighted in this study with the RoBERTa model.

5. Conclusion

In this study, we showed that classic supervised approaches were able, with promising results, to detect relevance and epidemiological information at the sentence level. Selected word embeddings improved the result of the classic bag-of-words representation for the classification of fine-grained epidemiological information. However, bag-of-words achieved comparable results for a binary classification task. Our results suggest that there is no turnkey solution for the choice of the textual representation, and the best model should be adapted to each classification task. Considering the classification performances obtained with minimal tuning of the word embedding model, we believe that further evaluation of the training parameters could enhance its quality. Several more specific questions about corpus size, preprocessing steps and classifier parameters tuning were not fully addressed and remain targets for future study. The size of the sliding window, for instance, has a marked effect on the vector similarities. Small windows tend to produce functional and syntactic similarities, while larger windows tend to produce more topical similarities (Goldberg, 2017). To evaluate these parameters, we aim to increase the annotated training corpus to override the constraints inherent to small training datasets in terms of evaluation robustness.

Funding

This work was funded by the French General Directorate for Food (DGAL), the French Agricultural Research Centre for International Development (CIRAD). This work was also supported by the French National Research Agency under the Investments for the Future Program, referred to as ANR-16-CONV-0004. This study was partially funded by EU grant 874850 MOOD and is catalogued as MOOD008. The contents of this publication are the sole responsibility of the authors and do not necessarily reflect the views of the European Commission.

Declaration of Competing Interest

The authors declare that they have no financial or personal interests that could influence the work reported in this paper.

References

Agibetov, A., Blagec, K., Xu, H., Samwald, M., 2018. Fast and scalable neural embedding models for biomedical sentence classification. BMC Bioinforma. 19 (1), 541. https://doi.org/10.1186/s12859-018-2496-4.

Arsevska, E., Valentin, S., Rabatel, J., de Goër de Hervé, J., Falala, S., Lancelot, R., Roche, M., 2018. Web monitoring of emerging animal infectious diseases integrated in the French Animal Health Epidemic Intelligence System. PLOS ONE 13 (8), e0199960. https://doi.org/10.1371/journal.pone.0199960.

Bahk, C.Y., Scales, D.A., Mekaru, S.R., Brownstein, J.S., Freifeld, C.C., 2015. Comparing timeliness, content, and disease severity of formal and informal source outbreak reporting. BMC Infect. Dis. 15 (1) https://doi.org/10.1186/s12879-015-0885-0.

Ben Jebara, K., Shimshony, A., 2006. International monitoring and surveillance of animal diseases using official and unofficial sources. Vet. Ital. 42 (4), 431–441.

S. Bird, E. Loper, NLTK: The Natural Language Toolkit, in: Proceedings of the ACL Interactive Poster and Demonstration Sessions, Association for Computational Linguistics, Barcelona, Spain, 2004, 214–217.⟨https://www.aclweb.org/anthology/P04–3031⟩.

Bollig, N., Clarke, L., Elsmo, E., Craven, M., 2020. Machine learning for syndromic surveillance using veterinary necropsy reports (publisher: Public Library of Science). PLOS ONE 15 (2), e0228105. https://doi.org/10.1371/journal.pone.0228105.

Brownlee, J., 2017. Deep learning for natural language processing: develop deep learning models for your natural language problems. Mach. Learn. Master (google-Books-ID: _pmoDwAAQBAJ).

Brownstein, J.S., Freifeld, C.C., Reis, B.Y., Mandl, K.D., 2008. Surveillance Sans Frontieres: Internet-based emerging infectious disease intelligence and the HealthMap project. PLoS Med. 5 (7), e151 https://doi.org/10.1371/journal.pmed.0050151.

D. Carter, M. Stojanovic, P. Hachey, K. Fournier, S. Rodier, Y. Wang, B. de Bruijn, Global Public Health Surveillance using Media Reports: Redesigning GPHIN, arXiv e-prints, 2020: arXiv:2004.04596_eprint: 2004.04596.

J. Chen, Y. Hu, J. Liu, Y. Xiao, H. Jiang, Deep short text classification with knowledge powered attention, in: Proceedings of the Thirty-Third AAAI Conference on and Thirty-First Innovative Applications of Conference and Ninth AAAI Symposium on Educational Advances in, AAAI'19/IAAI'19/EAAI'19, AAAI Press, 2019.10.1609/aaai.v33i01.33016252.

Conway, M., Doan, S., Kawazoe, A., Collier, N., 2009. Classifying Disease Outbreak Reports Using N-grams and Semantic. Int. J. Med. Inform. 78 (12).

C. d'Amato, M. Fernandez, V. Tamma, F. Lecue, P. Cudré-Mauroux, J. Sequeda, C. Lange, J. Heflin, The Semantic Web - ISWC 2017: 16th International Semantic Web Conference, Vienna, Austria, October 21–25, 2017, Proceedings, Part I, Springer, 2017, google-Books-ID: qHg5DwAAQBAJ.

De Boom, C., Van Canneyt, S., Demeester, T., Dhoedt, B., 2016. Representation learning for very short texts using weighted word embedding aggregation. Pattern Recognit. Lett. 80, 150–156. https://doi.org/10.1016/j.patrec.2016.06.012. 〈http://arxiv.org/abs/1607.00570〉.

J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019b, 4171–4186.10.18653/v1/N19–1423, 〈https://aclanthology.org/N19–1423〉.

J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv:1810.04805 [cs]ArXiv: 1810.04805.〈http://arxiv.org/abs/1810.04805〉 2019a.

Dion, M., AbdelMalik, P., Mawudeku, A., 2015. Big Data and the Global Public Health Intelligence Network (GPHIN). Can. Commun. Dis. Rep. 41 (9), 209–214. 〈https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5933838/〉.

Elgeldawi, E., Sayed, A., Galal, A.R., Zaki, A.M., 2021. yperparameter tuning for machine learning algorithms used for arabic sentiment analysis. Informatics 8 (4). https://doi.org/10.3390/informatics8040079. 〈https://www.mdpi.com/2227-9709/8/4/79〉.

S. Gella, L. DuongThanh, Automatic sentence classifier using sentence ordering features for event based medicine: Shared task system description, in: Proceedings of the Australasian Language Technology Association Workshop 2012, Dunedin, New Zealand, 2012, 130–133.〈https://aclanthology.org/U12–1018〉.

S. Ghosh, P. Chakraborty, E. Cohn, J.S. Brownstein, N. Ramakrishnan, Characterizing Diseases from Unstructured Text: A Vocabulary Driven Word2vec Approach, arXiv: 1603.00106 [cs, stat]ArXiv: 1603.00106.〈http://arxiv.org/abs/1603.00106〉 2016.

S. Ghosh, P. Chakraborty, B.L. Lewis, M.S. Majumder, E. Cohn, J.S. Brownstein, M.V. Marathe, N. Ramakrishnan, Guided Deep List: Automating the Generation of Epidemiological Line Lists from Open Sources, arXiv:1702.06663 [cs]ArXiv: 1702.06663.〈http://arxiv.org/abs/1702.06663〉 2017.

Go, A., Bhayani, R., Huang, L., 2009. Twitter sentiment classification using distant supervision. Processing 150.

Goldberg, Y., 2017. Neural Network Methods for Natural Language Processing. Synth. Lect. Hum. Lang. Technol. 10 (1), 1–309. https://doi.org/10.2200/S00762ED1V01Y201703HLT037.

Hartley, D., Nelson, N., Walters, R., Arthur, R., Yangarber, R., Madoff, L., Linge, J., Mawudeku, A., Collier, N., Brownstein, J., Thinus, G., Lightfoot, N., 2010. The landscape of international event-based biosurveillance. Emerg. Health Threats J. 3 (0) https://doi.org/10.3402/ehtj.v3i0.7096. 〈http://journals.co-action.net/index.php/ehtj/article/view/7096〉.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second edition.,. Springer Science & Business Media,. google-Books-ID: tVIjmNS3Ob8C.

M.A. Hearst, Untangling Text Data Mining, in: Proceedings of the 37th Annual Meeting of the Association for, Association for, College Park, Maryland, USA, 1999, 3–10.10.3115/1034678.1034679, 〈https://www.aclweb.org/anthology/P99–1001〉.

Heymann, D., Rodier, G., 2001. Hot spots in a wired world: WHO surveillance of emerging and re-emerging infectious diseases. Lancet Infect. Dis. 1 (5), 345–353.

Huang, C.-C., Lu, Z., 2015. Community challenges in biomedical text mining over 10 years: success, failure and the future. Brief. Bioinforma. 17 (1), 132–144. https://doi.org/10.1093/bib/bbv024. 〈http://arxiv.org/abs/https://academic.oup.com/bib/article-pdf/17/1/132/6685180/bbv024.pdf〉.

Joachims, T., 1998. Text categorization with Support Vector Machines: Learning with many relevant features. In: Carbonell, J.G., Siekmann, J., Goos, G., Hartmanis, J., van Leeuwen, J., Nédellec, C., Rouveirol, C. (Eds.), Machine Learning: ECML-98, Vol. 1398. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 137–142. https://doi.org/10.1007/BFb0026683.

A. Khoo, Y. Marom, D. Albrecht, Experiments with Sentence Classification, in: Proceedings of the Australasian Language Technology Workshop 2006, Sydney, Australia, 2006, 18–25.〈https://www.aclweb.org/anthology/U06–1005〉.

Kim, M., Chae, K., Lee, S., Jang, H.-J., Kim, S., 2020. Automated Classification of Online Sources for Infectious Disease Occurrences Using Machine-Learning-Based Natural Language Processing Approaches. Int. J. Environ. Res. Public Health 17 (24), E9467. https://doi.org/10.3390/ijerph17249467.

Krzeszewska, U., Poniszewska-Marańda, A., Ochelska-Mierzejewska, J., 2022. Systematic Comparison of Vectorization Methods in Classification Context. number: 10 Publisher: Multidisciplinary Digital Publishing Institute Appl. Sci. 12 (10), 5119. https://doi.org/10.3390/app12105119. number: 10 Publisher: Multidisciplinary Digital Publishing Institute. 〈https://www.mdpi.com/2076-3417/12/10/5119〉.

Kumar, M.A., Gopal, M., Comparison, A., 2010. Study on multiple binary-class SVM methods for unilabel text categorization. Pattern Recogn. Lett. 31 (11), 1437–1444. https://doi.org/10.1016/j.patrec.2010.02.015.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A Robustly Optimized BERT Pretraining Approach ArXiv: 1907.11692 [cs], 2019.10.48550/arXiv.1907.11692, 〈http://arxiv.org/abs/1907.11692〉.

Mandelbaum, A., Shalev, A., Word Embeddings and Their Use In Sentence Classification Tasks, 2016. arXiv:1610.08229 [cs]ArXiv: 1610.08229.〈http://arxiv.org/abs/1610.08229〉.

T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, Distributed Representations of Words and Phrases and Their Compositionality, in: Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2, NIPS'13, Curran Associates Inc., USA, 2013a, 3111–3119.〈http://dl.acm.org/citation.cfm?id=2999792.2999959〉.

T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient Estimation of Word Representations in Vector Space, 2013b. arXiv:1301.3781 [cs]ArXiv: 1301.3781.〈http://arxiv.org/abs/1301.3781〉.

T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, A. Joulin, Advances in pre-training distributed word representations, arXiv preprint arXiv:1712.09405, 2017.

Naili, M., Chaibi, A.H., Ben Ghezala, H.H., 2017. Comparative study of word embedding methods in topic segmentation. Procedia Comput. Sci. 112, 340–349. https://doi.org/10.1016/j.procs.2017.08.009. 〈https://linkinghub.elsevier.com/retrieve/pii/S1877050917313480〉.

Naughton, M., Stokes, N., Carthy, J., 2010. Sentence-level event classification in unstructured texts. Inf. Retr. 13 (2), 132–156. https://doi.org/10.1007/s10791-009-9113-0.

Paquet, C., Coulombier, D., Kaiser, R., Ciotti, M., 2006. Epidemic intelligence: a new framework for strengthening disease surveillance in Europe. Eurosurveillance 11 (12), 5–6. https://doi.org/10.2807/esm.11.12.00665-en.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al., 2011. Scikit-learn: Machine learning in python. J. Mach. Learn. Res. 12 (Oct), 2825–2830.

J. Pennington, R. Socher, C. Manning, Glove: Global Vectors for Word Representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar, 2014, 1532–1543.10.3115/v1/D14–1162, 〈http://aclweb.org/anthology/D14–1162〉.

J. Piskorski, J. Haneczok, G. Jacquet, New benchmark corpus and models for fine-grained event classification: To BERT or not to BERT?, in: Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Barcelona, Spain (Online), 2020, 6663–6678.10.18653/v1/2020.coling-main.584, 〈https://www.aclweb.org/anthology/2020.coling-main.584〉.

Pyysalo, S., Ohta, T., Rak, R., Rowley, A., Chun, H.-W., Jung, S.-J., Choi, S.-P., Tsujii, J., Ananiadou, S., 2015. Overview of the Cancer Genetics and Pathway Curation tasks of BioNLP Shared Task 2013. BMC Bioinforma. 16 (10), S2. https://doi.org/10.1186/1471-2105-16-S10-S2.

R. Rˆehůrˆ, P. Sojka Software Framework for Topic Modelling with Large Corpora, in: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, ELRA, Valletta, Malta, 2010, 45–50.

Salton, G., 1971. The SMART Retrieval System-Experiments in Automatic Document Processing. Prentice-Hall, Inc.,, USA.

Salton, G., Buckley, C., 1988. Term-weighting approaches in automatic text retrieval. Inf. Process. Manag. 24 (5), 513–523. https://doi.org/10.1016/0306-4573(88)90021-0. 〈http://www.sciencedirect.com/science/article/pii/0306457388900210〉.

Song, G., Ye, Y., Du, X., Huang, X., Bie, S., 2014. Short text classification: a survey. J. Multimed. 9 (5).

G. Soto, R.V. Araujo-Castillo, J. Neyra, M. Fernandez, C. Leturia, C.C. Mundaca, D.L. Blazes, Challenges in the implementation of an electronic surveillance system in a resource-limited setting: Alerta, in Peru, in: BMC proceedings, Vol. 2, BioMed Central, 2008, S4.

Thapen, N., Simmie, D., Hankin, C., 2016. The early bird catches the term: combining twitter and news data for event detection and situational awareness. J. Biomed. Semant. 7 (1), 61. https://doi.org/10.1186/s13326-016-0103-z.

Torregrossa, F., Allesiardo, R., Claveau, V., Kooli, N., Gravier, G., 2021. A survey on training and evaluation of word embeddings, 0-0 Int. J. Data Sci. Anal. 0 (0). https://doi.org/10.1007/s41060-021-00242-8.

Uysal, A.K., Gunal, S., 2014. The impact of preprocessing on text classification. Inf. Process. Manag. 50 (1), 104–112. https://doi.org/10.1016/j.ipm.2013.08.006. 〈http://www.sciencedirect.com/science/article/pii/S0306457313000964〉.

S. Valentin, R. Lancelot, M. Roche, Automated Processing of Multilingual Online News for the Monitoring of Animal Infectious Diseases, in: Proceedings of the LREC 2020 Workshop on Multilingual Biomedical Text Processing (MultilingualBIO 2020), European Language Resources Association, Marseille, France, 2020a, 33–36.〈https://www.aclweb.org/anthology/2020.multilingualbio-1.6〉.

S. Valentin, V. De Waele, A. Vilain, E. Arsevska, R. Lancelot, M. Roche, Annotation of epidemiological information in animal disease-related news articles: guidelines and manually labelled corpus, Dataverse CiradType: dataset.10.18167/DVN1/YGAKNB, 2019. 〈https://dataverse.cirad.fr/dataset.xhtml?persistentId=doi%3A10.18167%2FDVN1%2FYGAKNB&version=DRAFT〉.

Valentin, S., Arsevska, E., Falala, S., de Goër, J., Lancelot, R., Mercier, A., Rabatel, J., Roche, M., 2020b. PADI-web: A multilingual event-based surveillance system for monitoring animal infectious diseases. Comput. Electron. Agric. 169, 105163 https://doi.org/10.1016/j.compag.2019.105163. 〈http://www.sciencedirect.com/science/article/pii/S0168169919310646〉.

S. Valentin, E. Arsevska, A. Vilain, V.D. Waele, R. Lancelot, M. Roche, Annotation of epidemiological information in animal disease-related news articles: guidelines, 2021. arXiv:2101.06150.

Wilson, K., Brownstein, J.S., 2009. Early detection of disease outbreaks using the Internet. Can. Med. Assoc. J. 180 (8), 829–831. https://doi.org/10.1503/cmaj.090215.

Witten, I.H., Frank, E., Hall, M.A., Pal, C.J., 2016. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann,.

Y. Wu , M. Schuster , Z. Chen , Q.V. Le , M. Norouzi , W. Macherey , M. Krikun , Y. Cao , Q. Gao , K. Macherey , J. Klingner , A. Shah , M. Johnson , X. Liu , Łukasz. Kaiser , S. Gouws , Y. Kato , T. Kudo , H. Kazawa , K. Stevens , G. Kurian , N. Patil , W. Wang , C. Young , J. Smith , J. Riesa , A. Rudnick , O. Vinyals , G. Corrado , M. Hughes , J. Dean , oogle's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation, 2016. arXiv:1609.08144 [cs]ArXiv: 1609.08144.⟨http://arxiv.org/abs/1609.08144⟩.

Yin, Z., Shen, Y., 2018. On the Dimensionality of Word Embedding. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 31. Curran Associates, Inc., pp. 887–898. ⟨http://papers.nips.cc/paper/7368-on-the-dimensionality-of-word-embedding.pdf⟩

Y. Zhang, B. Liu, Semantic text classification of emergent disease reports, in: Proceedings of the 11th European Conference on Principles and Pratice of Knockledge Discovery in Databases (PKDD), Springer, Warsaw, Poland, 2007.

Zhang, Y., Jin, R., Zhou, Z., 2010. Understanding bag-of-words model: A statistical framework. Int. J. Mach. Learn. Cybern. 1 (1–4), 43–52. ⟨https://www.researchgate.net/publication/226525014_Understanding_bag-of-words_model_A_statistical_framework⟩.

Zhao, R., Mao, K., 2018. Fuzzy bag-of-words model for document representation (conference Name: IEEE Transactions on Fuzzy Systems). IEEE Trans. Fuzzy Syst. 26 (2), 794–804. https://doi.org/10.1109/TFUZZ.2017.2690222.