A Dataset for Multi-lingual Epidemiological Event Extraction

Stephen Mutuvi, Antoine Doucet, Gael Lejeune, Moses Odeo
Multimedia University - Kenya, La Rochelle University - France, Sorbonne University - France
smutuvi@mmu.ac.ke, antoine.doucet@univ-lr.fr, gael.lejeune@sorbonne-universite.fr, modeo@mmu.ac.ke

Abstract

This paper proposes a corpus for development and evaluation of tools and techniques for identifying emerging infectious disease threats in online news text. The corpus can not only be used for Information Extraction, but also for other Natural Language Processing tasks such as text classification. We make use of articles published on the Program for Monitoring Emerging Diseases (PROMED) platform, which provides current information about outbreaks of infectious disease globally. Among the key pieces of information present in the articles is the Uniform Resource Locator (URL) to the online news sources where the outbreaks were originally reported. We detail the procedure followed to build the dataset, which include leveraging the source URLs to retrieve the news reports and subsequently pre-processing the retrieved documents. We also report on experimental results of event extraction on the dataset using the Data Analysis for Information Extraction in any Language(DANIEL) system. DANIEL is a multilingual news surveillance system that leverages unique attributes associated with news reporting repetition and saliency, to extract events. The system has a wide geographical and language coverage, including low-resource languages. In addition, we compare different classification approaches in terms of their ability to differentiate between epidemic related and non-related news articles that constitute the corpus.

Keywords: Epidemiology, Corpus Creation, Event Extraction, Classification, Multingual

1. Introduction

Web corpora, a term that describes text corpora created from the web, have recently became popular due to the availability of a vast amount of electronic texts on the Web. The World Wide Web is a valuable source of data, which enable building of corpora with wide ranging attributes such as varying sizes, languages and domains. Such corpora can be analysed and utilized in key application areas, among them epidemic surveillance.

Epidemic intelligence is an integral component of infectious diseases early-warning mechanisms. It involves collection, analysis and dissemination of key information related to disease outbreaks, with an objective of detecting outbreaks and providing early warning to public health stakeholders (World Health Organization, 2014). Disease surveillance mechanisms can broadly be classified as either indicator- or event-based surveillance (Huff et al., 2016).

Indicator-based surveillance are the conventional surveillance systems which rely predominantly on local health practitioners to identify infectious disease outbreaks, where suspected outbreak cases are subjected to laboratory tests for confirmation. A key determinant of efficiency of such surveillance methods is the underlying health care infrastructure. Poor infrastructure can result to inaccurate and irrelevant information being disseminated, with a likelihood of significant time delays in dissemination of key information relevant to disease outbreaks (Zhou et al., 2011). Inadequate health infrastructure directly often leads to incomplete geographical coverage when reporting epidemics (Huff et al., 2016; World Health Organization, 2014). Time delays and incomplete coverage may hinder deployment of effective health interventions, potentially leading to loss of lives.

Today, using NLP to monitor informal information sources, such as social media, search queries, online news outlets and blogs has become an essential part of epidemic surveillance (Salathé et al., 2013; Bernardo et al., 2013). Advancements in NLP presents an opportunity to efficiently collect, process and analyze large textual data from the web,to detect disease related features from the text. Near real-time data-driven surveillance systems, commonly referred to as Event-Based Surveillance (EBS) systems can now be easily developed and deployed in production. Event-Based Surveillance encompasses analysing textual data mostly generated via the web, for incidences of or events related to disease outbreaks (Huff et al., 2016). A more plausible approach is utilizing a combination of formal and informal sources for timely and accurate detection infectious disease outbreak (O'Shea, 2017).As such, it has been determined that event-based surveillance methods can complements the traditional surveillance methods, for timely and accurate detection of epidemics (Chunara et al., 2012; O'Shea, 2017).

However, despite the rise in use of advanced text processing and analysis, such as deep learning, there exists limited number of corpus for training and evaluation of disease events extraction models. The few available datasets are relatively small in size and predominantly in English language. A key requirement for training deep learning models that give satisfactory results is having sufficiently large-scale datasets. This is further compounded by the fact that epidemic reports originate from a wide rage of sources and languages.

In view of the above, we attempt to address the dearth of data for epidemic event extraction, by creating a corpus that can be used by researchers and practitioners in building and evaluate epidemic event extraction algorithms and applications. We leverage the Program for Monitoring Emerging Diseases (PROMED) reporting platform to create the corpus. PROMED aggregates disease outbreak reports across the world and is open and publicly available. The PROMED articles undergo a review and verification process by experts before being published on the platform. The aggregation of the reports by subject matter experts

make the articles suitable for use as ground truth to evaluate epidemiological information extraction systems. The multilingual dataset we extracted from PROMED comprises articles in English, French, Portuguese and Spanish languages. To the best of our knowledge, this is among the largest datasets of this nature that is available for developing and evaluating multilingual epidemic surveillance tools and techniques.

The paper is organized as follows. Section 2. reviews related work on event extraction while Section 3. describes the methodology used to create the corpus. Section 4. describes experiments to train the corpus in a text classification task. Additionally, we evaluate event extraction over the corpus using the DANIEL system. The results are discussed in Section 5. before conclusions are drawn and future work presented in Section 6..

2. Related Work

Event extraction (EE) is an important information extraction (IE) task that focuses on identifying an event mention from text and extracting information relevant to the event. Typically, this entails predicting event triggers, the occurrence of events with specific types, and extracting arguments associated with an event.

While event extraction is a crucial sub-task of information extraction, it still remains quite a challenging task due to the difficult associated with encoding words semantics in various context (Zhan and Jiang, 2019). For instance, same event might appear in the form of various trigger expressions or an expression might represent different event types in different contexts.

Event extraction methods are classified into three, namely pattern-based, data-driven and hybrid methods (Hogenboom et al., 2011). Pattern-based methods employ use of rules and templates to extract events from text through representation and exploitation of expert knowledge. On the other hand, data-driven approaches use statistical techniques to discover the relations in text. An approach that combines rule-based and data-driven methods is referred to as an hybrid approach. The methods based on rules and templates are more mature, the methods based on statistical machine learning are dominant while method based on deep learning have recently gained popularity among researchers in the field (Zhan and Jiang, 2019).

Specific to epidemiological event extraction, there exist a number of empirical works targeted to extractions of events related to disease outbreaks. Among them is Data Analysis for Information Extraction in any Language (DANIEL), a multilingual news surveillance system that leverages repetition and saliency, properties that are common in news writing (Lejeune et al., 2015). The multilingual nature of the system enable global and timely detection of epidemic events since it eliminates the requirement for translating local news to other languages for subsequent transmission. The system can easily be adapted and scaled to extract events across languages, therefore, being able to have a wider geographical coverage. Reactivity and geographic coverage are of paramount importance in epidemic surveillance (Lejeune et al., 2015).

Similar to DANIEL are BIOCASTER (Collier, 2011; Collier et al., 2008) and PULS (Du et al., 2011) which have produced good results in analysing disease-related news reports and providing a summary of the epidemics. The EcoHealth Alliance Global Rapid developed the Identification Tool System (GRITS), an application that provides automatic analyses of epidemiological texts. The system extracts important information about a disease outbreak, such as the most likely disease, dates and countries where the outbreak originates. The pipeline for GRIT entails transforming words to vectors using TF-IDF, extracting features using pattern-matching tools, before applying binary relevance-based classifier to predict the available disease in the text (Huff et al., 2016).

Internet search data has also been exploited for disease surveillance. In one study, internet searches for specific cancers was found to correlate with their estimated incidence and mortality (Cooper et al., 2005). Monitoring influenza outbreak using data drawn from the web has also been previously explored. Two different studies, one utilizing GOOGLE (Ginsberg et al., 2009) and the other YAHOO (Polgreen et al., 2008) search queries, analysed the searches and estimated the number of reported Influenza cases. In the recent years, a flurry of work has utilized social media data for infectious disease surveillance (Paul et al., 2016; Charles-Smith et al., 2015). Mostly, Twitter data, has been used for disease tracking (Lamb et al., 2013; Collier et al., 2011; Culotta, 2010), outbreak detection (Li and Cardie, 2013; Bodnar and Salathé, 2013; Diaz-Aviles et al., 2012; Aramaki et al., 2011) and predicting the likelihood of individuals falling sick (Sadilek et al., 2012). News media has also been used to give early warning of increased disease activity before official sources have reported (Brownstein et al., 2008). The studies have demonstrated the potential value of harnessing data-driven approaches for epidemic surveillance.

3. Methods

In this section, we describe the procedure followed to create the corpus. We also detail the process for evaluating event extraction and classification models over the corpus.

3.1. Corpus Creation

We retrieved PROMED articles in English, French, Spanish and Portuguese languages, for the period August, 1, 2013 to August, 31, 2019. The articles contained various key meta-data such as title, description, location, date and source URL where the article was originally published. The source URLs present in the PROMED articles were extracted and their corresponding source documents downloaded. Figure 1 shows the percentage of documents still available online for each year in the date range 01-08-2013 to 31-08-2019. The source URLs, together with the other meta-data were formatted and stored in json format making corpus[1] easily reusable and reproducible. Therefore, this makes it easy for any interested researcher to process the dataset and use it in modeling epidemiological event extraction or any other related NLP tasks.

[1] https://github.com/smutuvi/epidemic_surveillance_corpus

Various processing tasks were performed on the extracted web data to transform the data into a clean text corpora. Firstly, language filtering was performed to ensure that only documents belonging to the languages of interest were retained. The documents were grouped into different clusters using Kmeans clustering algorithm. This enabled identification and separation of documents with content from blank documents. The silhouette coefficient was computed to quantify the appropriate number of clusters for each set of data. This coefficient measures how well data are assigned to its own cluster and how far they are from other clusters (Rousseeuw, 1987). A coefficient close to 1 (one) means the data sample is located in the appropriate cluster while -1 (negative one) implies data has been assigned to the wrong cluster. Elimination of boilerplate content from the corpus was among the data cleaning tasks. Content such as navigation links, headers and footers being removed from the from HTML pages using the JUSTEXT library (Pomikálek, 2011). Removal of boilerplate content is highly desirable, since such content rarely provide useful evidence about the phenomenon being investigated. On the contrary, the high frequency of the boilerplate content could introduce bias into the text data, hence negatively impacting the performance of derived applications (Vogels et al., 2018). The final pre-processing task was deduplication. Deduplication involves eliminating perfect duplicate and near-duplicate content so that only one instance of each text was preserved. The ONION (ONe Instance ONly tool (Pomikálek, 2011), which deduplicate text data by measuring the similarity of paragraphs or entire document was used. Onion is based on a n-gram-based one-pass deduplication algorithm, where for each document all n-grams of words are extracted (10-grams by default) and compared with the set of previously seen n-grams (Pomikálek, 2011).

3.2. Non-Epidemic Dataset

In addition to the disease outbreak-related dataset, we also prepared a non-related dataset for training a text classification model. The dataset constituted news articles from the News Category Dataset (Misra, 2018), consisting of around 200,000 English news articles. The news articles, which do not have mentions outbreak of diseases, were published on the huffpost news website between the year 2012 and 2018. The dataset categorize news articles based on their headlines and short descriptions. The news articles are grouped into various categories such as politics, wellness, travel, entertainment, sports, healthy living among others. A total of 7,325 articles corresponding to politics, entertainment and sports categories were selected and downloaded from HuffPost news platform. Similar to ProMED articles, language filtering, cleaning boilerplate content, clustering to help identify empty documents and deduplication were undertaken to clean and process the corpus.

3.3. Corpus Statistics

The corpus statistics, PROMED and source documents, is presented on Table 1 and Table 2 respectively.

Table 3 presents statistics for the corpus used for training and evaluating text classification models. The dataset is composed of epidemic relevant articles from ProMED and non-relevant documents from News Category Dataset, described in Section 3.2. A total of 10,000 and 2,996 documents in English and French language, comprising relevant and non-relevant documents formed the training set. The relevant and non-relevant documents were equally distributed among the two classes.

Human-annotated datasets (Lejeune et al., 2015) in English and French provided the groundtruth to evaluate the models. The availability of the annotated dataset for the two languages informed the decision for their consideration in our experiments. The two other language,Spanish and Portuguese, which currently do not have groundtruth data will be considered in our future studies. The annotated datasets comprising 444 and 2,722 documents for English and French languages respectively. The test data had high degree of imbalance between the classes. The 444 English test documents had 31 relevant and 413 non-relevant documents. The French test corpus comprised 299 relevant and 2,207 non-relevant documents. The imbalance was important in depicting the models' ability to classify the documents into their respective classes.

3.4. Evaluation

We describe the procedure for extracting epidemic events present in the corpus using DANIEL. The performance of

Language	#Documents	#Sentences	#Words
English (en)	19,149	558,448	53,325,455
French (fr)	1,849	28,823	5,593,184
Spanish (es)	3,453	27,918	4,458,533
Portuguese (pt)	3,451	48,591	5,994,583

Table 1: Statistics for Retrieved PROMED Documents

Language	#Documents	#Sentences	#Words
English	13,275	320,613	8,749,272
French	1,395	13,777	439,153
Spanish	1,994	27,751	863,672
Portuguese	1,562	14,424	528,701

Table 2: Statistics for Retrieved Source Documents

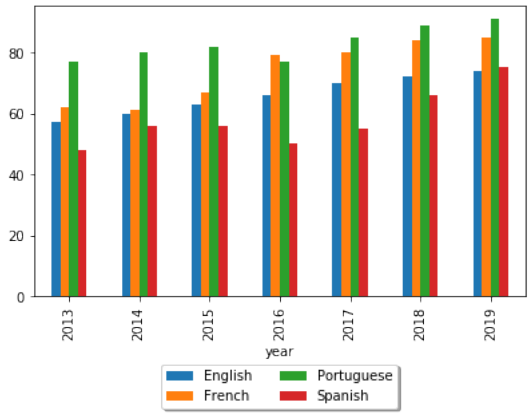

Figure 1: Percentage of ProMED sources accessible by year.

Dataset	#Documents	#Sentences	#Words
Train-en	10,000	317,862	9,879,559
Train-fr	2,996	43,257	1,959,584
Test-en	444	4,728	230,353
Test-fr	2,722	75,479	2,058,941

Table 3: Statistics for train and test datasets used in training and evaluating the classification models

Naive Bayes, Random Forest and Neural Network classification models in classifying documents as either epidemic-related or not is evaluated.

3.4.1. The DANIEL System

We evaluated the performance of the DANIEL system in extracting epidemic events present in the corpus. The DANIEL processing pipeline comprises three steps: news article segmentation, event detection and event localization. DANIEL adopts a discourse-level event extraction approach, where the global structure of news is exploited (Lejeune et al., 2015). The system relies on properties that are common to the journalistic genre regardless of the language. The most useful features are repetition and saliency, which defines the relative importance of prominence of news contents. While majority of systems extract events at sentence level, by harnessing the morphological, syntactical and semantic features of a sentence, hence dependent on language specific modules, DANIEL uses language agnostic text level features. It is character-based, hence handles text as a sequence of characters rather than as a sequence of words. Rather than exploiting keywords, the system exploits strings of text, but only if the strings have been repeated in pre-defined salient zones in text. The output of the DANIEL system is a disease-location pair describing an event as a disease outbreak and the place where it occurred. Recall and precision scores were obtained to determined the performance. The results are presented in Section 4..

For further evaluation, subsets of the English and French language datasets were subjected to annotation by 3 native speakers for each language. These annotators had to judge whether documents presented to them had mentions of infectious disease outbreak or not. Subsequently, for the relevant documents, the annotators were requested to specify the disease name and location. We measured the Inter-rater reliability using Cohen's kappa coefficient. Inter-rater reliability determines the extent to which data collectors (raters) assign the same score to the same variable (McHugh, 2012). Finally, leveraging the generated ground-truth, evaluation was quantitatively done against annotators judgements on the evaluation corpus.

3.4.2. Text Classification Models

We train and evaluate text classification models using datasets described in Table 3. The models classify a news article as either relevant or non-relevant, depending on whether it alerts about a disease outbreak or not. The training data comprised 10,000 and 2,722 news articles in English and French languages respectively, with documents equally distributed among the two classes. Pre-processing of the text input was undertaken which included cleaning and tokenizing the data. Text preprocessing, tokenizing and filtering of stopwords functions are all included in TfidfVectorizer class of scikit-learn library (Pedregosa et al., 2011). Additionally, the TfidfVectorizer class facilitates feature extraction by enabling the generation of a dictionary of features and by transforms documents to feature vectors. The learned vocabulary can be used to encode new documents. A human annotated dataset presented in Table 3 was used as the test set to evaluate the performance of the classification models. With the data ready, we trained a Multinomial Naive Bayes, Random Forest and Neural Network classifiers using the created corpus. Naive bayes classifier was used as the base model. Naive Bayes has been proven to be viable for text classification and information retrieval in general (Le et al., 2019). Parameter tuning and transfer learning were undertaken for the Random Forest and Neural-based classifiers respectively, with a goal of enhancing their performance. Finally, the models were evaluated to determine their performance using the human annotated test data. We evaluated the models' performance by comparing the actual and predicted labels. In our case, classification accuracy could not suffice because the two classes were not balanced. Evaluation metrics such as recall, precision and F-measure are more appropriate if there existed a greater degree of imbalance in the classes (Bunker and Thabtah, 2017).

4. Results

In this section, the results of event extraction and text classification models are presented.

4.1. Event Extraction

This section presents the results of event extraction on the created text corpus using DANIEL, an epidemiology event extraction system. The measures used for evaluation were recall, precision and F-measure.The measures are briefly described as follows:

- Recall is the proportion of relevant items identified correctly: Recall = $\frac{TruePositive}{Truepositives+FalseNegatives}$
- Precision is the proportion of retrieved items that are relevant: Precision = $\frac{TruePositive}{TruePositive+FalsePositive}$
- F-measure is the harmonic mean of Precision and Recall: F-measure = 2* $\frac{(Recall*Precision)}{(Recall+Precision)}$

DANIEL system attained a F-score of 75% for documents in both English and French languages. For the documents in English language, the system achieved a precision of 60% and was able to correctly identify all the relevant documents. Precision and recall scores of 74% and 83% were obtained on the French language documents.

4.2. Text Classification

The results for text classification models trained and evaluated using the datasets built in this study are presented. The models' F-score on the English test data was 74%, 63% and 53% for Random Forest, Neural Network and Naive Bayes model respectively. For the French document, a F-score of 67%, 63% and 50% was obtained for Random Forest,

Naive Bayes and Neural Network model respectively. The precision, recall and F-measure values for the three models, using the English and French language corpus are presented in Table 4 and Table 5 respectively.

Classifier	Precision	Recall	F-measure
Naive Bayes	57%	75%	53%
Random Forest	80%	70%	74%
Neural Network	68%	76%	61%

Table 4: Text Classification Report for the English Documents

Classifier	Precision	Recall	F-measure
Naive Bayes	62%	74%	63%
Random Forest	80%	63%	67%
Neural Network	64%	52%	50%

Table 5: Text Classification Report for the French Documents

5. Discussion

Considering the F-measure, the event extraction system's performance was generally good for both languages. However, it was noted that the system's ability to detect relevant documents correctly was better for English documents compared to the French language documents. This can be attributed to the fact that Daniel's rule-based inference engine leverages language and disease text resources, which are readily available for English language compared to other languages.

For the classification task, the Random Forest model had the highest F-measure compared to the other models. However, for French documents, the model could predict only one class using the default classification threshold of 0.5. This necessitated experimenting with a lower threshold of 0.3, which produced superior results compared to the other models. We conjecture that the Random Forest model produced the best results due to the fine-tuning of the model's parameters. However, the process of tuning the parameters required fairly significant effort and time. The performance of the neural network model was equally good. This can be attributed to their ability to automatically learn discriminating and reliable features from the text corpus. The performance can also be attributed to the use of transfer learning via pre-trained language models, specifically the Bidirectional Encoder Representations from Transformers (BERT). Such language models enable learning of contextualized representations which upon fine-tuning for tasks such as classification usually results in significant performance gains. Typically, language models are trained on large text corpora, hence being able to adequately capture linguistic features and representations, which results to the improved performance of downstream tasks.

6. Conclusion

Early detection of disease outbreaks is critical for deployment of effectively public health interventions. Delayed interventions may result to unprecedented calamity, which could include loss of lives. In addition to reactivity, coverage of epidemiological event detection systems is equally of paramount importance, particularly because outbreaks are reported from different parts of the world in different languages. Taking this into account, computation approaches, referred in this paper as event-based surveillance systems suffice. A key requirement for development of such systems is large multi-lingual datasets to train and evaluate high performance machine learning models. It's evident from existing literature that such large datasets are not adequate particularly for epidemiological surveillance settings. In this study, we attempt to contribute towards solving this challenge by developing and making available a large multi-lingual dataset suitable for training and evaluation of epidemiological event extraction models. The dataset can also be used for other NLP tasks such as text classification, text summarization among others.

7. Acknowledgments

This work has been supported by the European Union Horizon 2020 research and innovation programme under grants 825153 (Embeddia) and 770299 (NewsEye).

8. Bibliographical References

Aramaki, E., Maskawa, S., and Morita, M. (2011). Twitter catches the flu: detecting influenza epidemics using twitter. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1568–1576. Association for Computational Linguistics.

Bernardo, T. M., Rajic, A., Young, I., Robiadek, K., Pham, M. T., and Funk, J. A. (2013). Scoping review on search queries and social media for disease surveillance: a chronology of innovation. *Journal of medical Internet research*, 15(7):e147.

Bodnar, T. and Salathé, M. (2013). Validating models for disease detection using twitter. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 699–702. Acm.

Brownstein, J. S., Freifeld, C. C., Reis, B. Y., and Mandl, K. D. (2008). Surveillance sans frontieres: Internet-based emerging infectious disease intelligence and the healthmap project. *PLoS medicine*, 5(7):e151.

Bunker, R. P. and Thabtah, F. (2017). A machine learning framework for sport result prediction. *Applied computing and informatics*.

Charles-Smith, L. E., Reynolds, T. L., Cameron, M. A., Conway, M., Lau, E. H., Olsen, J. M., Pavlin, J. A., Shigematsu, M., Streichert, L. C., Suda, K. J., et al. (2015). Using social media for actionable disease surveillance and outbreak management: a systematic literature review. *PloS one*, 10(10):e0139701.

Chunara, R., Freifeld, C. C., and Brownstein, J. S. (2012). New technologies for reporting real-time emergent infections. *Parasitology*, 139(14):1843–1851.

Collier, N., Doan, S., Kawazoe, A., Goodwin, R. M., Conway, M., Tateno, Y., Ngo, Q.-H., Dien, D., Kawtrakul, A., Takeuchi, K., et al. (2008). Biocaster: detecting public health rumors with a web-based text mining system. *Bioinformatics*, 24(24):2940–2941.

Collier, N., Son, N. T., and Nguyen, N. M. (2011). Omg u got flu? analysis of shared health messages for bio-surveillance. *Journal of biomedical semantics*, 2(5):S9.

Collier, N. (2011). Towards cross-lingual alerting for bursty epidemic events. *Journal of Biomedical Semantics*, 2(5):S10.

Cooper, C. P., Mallon, K. P., Leadbetter, S., Pollack, L. A., and Peipins, L. A. (2005). Cancer internet search activity on a major search engine, united states 2001-2003. *Journal of medical Internet research*, 7(3):e36.

Culotta, A. (2010). Detecting influenza outbreaks by analyzing twitter messages. *arXiv preprint arXiv:1007.4748*.

Diaz-Aviles, E., Stewart, A., Velasco, E., Denecke, K., and Nejdl, W. (2012). Epidemic intelligence for the crowd, by the crowd. In *Sixth International AAAI Conference on Weblogs and Social Media*.

Du, M., Von Etter, P., Kopotev, M., Novikov, M., Tarbeeva, N., and Yangarber, R. (2011). Building support tools for russian-language information extraction. In *International Conference on Text, Speech and Dialogue*, pages 380–387. Springer.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., and Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012.

Hogenboom, F., Frasincar, F., Kaymak, U., and De Jong, F. (2011). An overview of event extraction from text. In *DeRiVE@ ISWC*, pages 48–57. Citeseer.

Huff, A. G., Breit, N., Allen, T., Whiting, K., and Kiley, C. (2016). Evaluation and verification of the global rapid identification of threats system for infectious diseases in textual data sources. *Interdisciplinary perspectives on infectious diseases*, 2016.

Lamb, A., Paul, M. J., and Dredze, M. (2013). Separating fact from fear: Tracking flu infections on twitter. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 789–795.

Le, C.-C., Prasad, P., Alsadoon, A., Pham, L., and Elchouemi, A. (2019). Text classification: Naïve bayes classifier with sentiment lexicon. *IAENG International Journal of Computer Science*, 46(2):141–148.

Lejeune, G., Brixtel, R., Doucet, A., and Lucas, N. (2015). Multilingual event extraction for epidemic detection. *Artificial intelligence in medicine*, 65(2):131–143.

Li, J. and Cardie, C. (2013). Early stage influenza detection from twitter. *arXiv preprint arXiv:1309.7340*.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica: Biochemia medica*, 22(3):276–282.

Misra, R. (2018). News category dataset, 06.

O'Shea, J. (2017). Digital disease detection: A systematic review of event-based internet biosurveillance systems. *International journal of medical informatics*, 101:15–22.

Paul, M. J., Sarker, A., Brownstein, J. S., Nikfarjam, A., Scotch, M., Smith, K. L., and Gonzalez, G. (2016). Social media mining for public health monitoring and surveillance. In *Biocomputing 2016: Proceedings of the Pacific symposium*, pages 468–479. World Scientific.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Polgreen, P. M., Chen, Y., Pennock, D. M., Nelson, F. D., and Weinstein, R. A. (2008). Using internet searches for influenza surveillance. *Clinical infectious diseases*, 47(11):1443–1448.

Pomikálek, J. (2011). *Removing boilerplate and duplicate content from web corpora*. Ph.D. thesis, Masarykova univerzita, Fakulta informatiky.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Sadilek, A., Kautz, H., and Silenzio, V. (2012). Predicting disease transmission from geo-tagged micro-blog data. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*.

Salathé, M., Freifeld, C. C., Mekaru, S. R., Tomasulo, A. F., and Brownstein, J. S. (2013). Influenza a (h7n9) and the importance of digital epidemiology. *The New England journal of medicine*, 369(5):401.

Vogels, T., Ganea, O.-E., and Eickhoff, C. (2018). Web2text: Deep structured boilerplate removal. In *European Conference on Information Retrieval*, pages 167–179. Springer.

World Health Organization. (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.

Zhan, L. and Jiang, X. (2019). Survey on event extraction technology in information extraction research area. In *2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, pages 2121–2126. IEEE.

Zhou, X., Ye, J., and Feng, Y. (2011). Tuberculosis surveillance by analyzing google trends. *IEEE transactions on biomedical engineering*, 58(8):2247–2254.