Multilingual Epidemiological Text Classification: A Comparative Study

Stephen Mutuvi
Multimedia University
Kenya
smutuvi@mmu.ac.ke

Emanuela Boros
University of La Rochelle
France
emanuela.boros@univ-lr.fr

Antoine Doucet
University of La Rochelle
France
antoine.doucet@univ-lr.fr

Gaël Lejeune
Sorbonne University
France
gael.lejeune@sorbonne-universite.fr

Adam Jatowt
Kyoto University
Japan
jatowt@gmail.com

Moses Odeo
Multimedia University
Kenya
modeo@mmu.ac.ke

Abstract

In this paper, we approach the multilingual text classification task in the context of the epidemiological field. Multilingual text classification models tend to perform differently across different languages (low- or high-resource), more particularly when the dataset is highly imbalanced, which is the case for epidemiological datasets. We conduct a comparative study of different machine and deep learning text classification models using a dataset comprising news articles related to epidemic outbreaks from six languages, four low-resourced and two high-resourced, in order to analyze the influence of the nature of the language, the structure of the document, and the size of the data. Our findings indicate that the performance of the models based on fine-tuned language models exceeds by more than 50% the chosen baseline models that include a specialized epidemiological news surveillance system and several machine learning models. Also, low-resource languages are highly influenced not only by the typology of the languages on which the models have been pre-trained or/and fine-tuned but also by their size. Furthermore, we discover that the beginning and the end of documents provide the most salient features for this task and, as expected, the performance of the models was proportionate to the training data size.

1 Introduction

Monitoring and containment of infectious disease outbreaks have been an ongoing challenge globally. Whether previously with Ebola or today with the Covid-19 pandemic, surveillance has remained a key component of public health strategy to contain the diseases. The ability to detect disease outbreaks in an accurate and timely manner is critical in the deployment of efficient intervention measures. For instance, Ebola cases and outbreaks need to be immediately detected in order to be contained and stopped. A delayed response can have a significant economic and social impact, in addition to increased morbidity and mortality rates. Thus, the detection needs to be done as soon as the first reports appear, and, naturally, as such reports may not be in English, there is a need for effective multilingual surveillance systems.

In recent years, there has been a rapid increase in data generated as a result of the progressive evolution of the Internet. The proliferation of digital data sources provide an avenue for data-driven surveillance, referred to as Epidemic Intelligence. Epidemic intelligence involves the collection, analysis, and dissemination of key information related to disease outbreaks, with the objective of detecting outbreaks and providing early warning to public health stakeholders (World Health Organization, 2014). Natural Language Processing (NLP) techniques have made it possible to analyze data from web sources, such as social media, search queries, blogs, and online news articles for health-related incidents and/or events (Salathé et al., 2013; Bernardo et al., 2013). Data-driven epidemic intelligence can be viewed as a two-step process comprising a classification task followed by the event extraction task (Joshi et al., 2019), which can help to predict the epidemic disease dynamics and where the next outbreak of epidemic would most likely happen. The classification task entails the identification of texts relevant to disease outbreaks

Proceedings of the 28th International Conference on Computational Linguistics, pages 6172–6183
Barcelona, Spain (Online), December 8-13, 2020

from a large collection of data. Considering, for instance, a dataset of online news articles, the articles that report an outbreak of disease are separated from those which do not. Next, the event extraction task takes the identified relevant documents as input and predicts health-related events with arguments such as the disease name and the location where the outbreak was reported. Like any other task that involves text analysis using NLP approaches, ambiguity is a key challenge when dealing with epidemic-related text data. Ambiguity manifests itself where a sentence in a document may have mentioned a disease, but may not necessarily be reporting on an outbreak of a disease. For instance, with the ongoing coronavirus pandemic, there are numerous news articles posted daily reporting on various aspects related to the disease. It becomes a challenge to extract the few relevant news articles that are of interest to the epidemiologist, articles that report on the number and location of new cases. Epidemic reporting is also characterized by news reports from divergent sources and languages which further compounds computational epidemiology. Furthermore, when working in a multilingual setting of real-world data, another challenge arises from the lack of annotated data for low-resource languages. The creation of such data can be expensive, time-consuming, and requires human expertise to annotate, hence it is a labor-intensive task.

In view of these challenges, appropriate NLP approaches are required in order for data-driven epidemic surveillance to be successful. Therefore, we seek to provide a comprehensive quantitative study of low-shot text classification models applied on a dataset comprising news articles about disease outbreaks from several diverse language families namely, English, Greek, French, Russian, Polish, and Chinese.

We seek to compare state-of-the-art approaches for epidemiological text classification from both deep learning and classical machine learning techniques by training a variety of models and evaluating them in several circumstances, in order to analyze their application in a real-world scenario. To the best of our knowledge, this is the first extensive study to specifically evaluate the performance of multilingual epidemiological text classification methods.

The remainder of this paper is organized as follows. Section 2 reviews works related to NLP-based epidemic surveillance systems, Section 3 describes the dataset used in the study, while the experiment setup and results are presented in Section 4. Finally, we provide a discussion of the results in Section 5, and the conclusions and possible suggestions for future research are presented in Section 6.

2 Related Work

There are a number of empirical works targeted at the application of NLP for the detection of disease outbreaks. Among them is Data Analysis for Information Extraction in any Language (DANIEL), a multilingual news surveillance system that leverages repetition and saliency (the beginning and the end of a news text often comprises the salient zones), properties that are common in news writing (Lejeune et al., 2015). By avoiding grammar analysis and the usage of language-specific NLP toolkits (e.g., Part-of-speech tagger, dependency parser) and by focusing on the general structure of journalistic writing style (Hamborg et al., 2018; Lucas, 2009), the system is able to detect crucial information from news articles. Furthermore, the multilingual nature of the system enables 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 is BIOCASTER (Collier, 2011; Collier et al., 2008) which has produced good results in analyzing disease-related news reports and in providing a summary of the epidemics. The BIOCASTER, an ontology-based text mining system, processes, and analyzes web text for the occurrence of disease outbreak in four phases namely, topic classification, Named Entity Recognition (NER), disease/location detection, and event recognition. The Naïve Bayes algorithm is used for the classification of the reports for topical relevance. The news stories identified to be relevant to disease outbreaks are propagated to the subsequent levels of processing. The major limitation of the BIOCASTER is its inability to detect disease outbreaks beyond the eight languages (Chinese, English, French, Japanese, Korean, Spanish, Thai, and Vietnamese) present in its ontology (Doan et al., 2019). Therefore, scaling the system to work across different languages requires manually updating the ontology with information for new languages. In addition, the system is not publicly available, except for the ontology.

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 GRITS entails transforming words to vectors using the term frequency-inverse document frequency (TF-IDF) method, by first extracting features using pattern-matching tools, before applying a binary relevance-based classifier to predict the presence of a disease name in the text (Huff et al., 2016). The system translates non-English documents using the Bing translator, which can potentially introduce errors to subsequent analysis steps if the translation is incorrect (Huff et al., 2016).

Internet search queries have also been exploited for disease surveillance. In one study, internet searches for specific cancers were found to be correlated 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 that analyzes large numbers of Google search queries to track influenza-like illness in a population (Ginsberg et al., 2009) and the other that examines search queries from Yahoo[1] related to the same aforementioned infectious disease (Polgreen et al., 2008) were conducted in this context.

In recent years, various studies have 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.

While prior attempts to develop multilingual epidemic surveillance systems have been made, the proposed systems are predominantly ontology-based, which require the ontologies to be updated on an ongoing basis in order to improve their performance and ensure broad coverage of different languages. Recently, there has been a growing interest in exploring the multilingual nature of the data in different domains, which could also be beneficial to the epidemiological domain. Existing multilingual methods use word representations that are either learned jointly using parallel corpora (Gouws et al., 2015; Luong et al., 2015) or via mapping separately trained word embeddings in different languages to a shared space through linear transformations (Artetxe et al., 2018; Mikolov et al., 2013). The embedding spaces of the different languages ought to have a similar structure for the linear mapping from one space to the other to be effective.

More recently, effective cross-lingual representations have been developed, by simultaneously training contextual word embedding models over multiple languages, without requiring mapping to a shared space. Such models learn representations of unlabeled data that generalize across languages. Examples include the cross-lingual language model (XLM) (Lample and Conneau, 2019) and multilingual BERT (Devlin et al., 2019) which are pre-trained on Wikipedia data for different languages. BERT has been shown to allow effective cross-lingual transfer on different downstream tasks. This includes, document classification (Qin et al., 2020; Wu and Dredze, 2019), named entity recognition (NER) (Wu and Dredze, 2019; Pires et al., 2019), sentiment classification (Qin et al., 2020), neural machine translation, (Kudugunta et al., 2019), and dependency parsing (Kondratyuk and Straka, 2019).

3 Dataset

We extend the dataset proposed by Mutuvi et al. (2020) to include additional languages so that it covers news articles from several, diverse language families: Germanic: English (en), Hellenic: Greek (el), Romance: French (fr), Slavic: Russian and Polish, and Chinese that descends from the Sino-Tibetan family. The articles were obtained from different online news sources with articles relevant to disease outbreak being obtained mainly via the Program for Monitoring Emerging Disease (ProMED)[2] platform,

[1] `http://search.yahoo.com`

[2] `https://promedmail.org/`

which is a program from the International Society for Infectious Diseases that tracks infectious disease outbreaks and acute exposures to toxins, across the world.

Language	#Documents	#Sentences	#Tokens
English (en)	3,562	117,190	2,692,942
French (fr)	2,415	70,893	1,959,848
Polish (pl)	341	9,527	151,901
Russian (ru)	426	6,865	133,905
Chinese (zh)	446	4,555	236,707
Greek (el)	384	6,840	183,373

Table 1: Dataset statistics.

The process of gathering the data involved first, retrieval of ProMED news articles published between August 1, 2013, and August 31, 2019. The articles clearly annotate the title, the description that captures details about the reported disease, location, date, and the source Uniform Resource Locator (URL) where the article was originally published. The source URLs were extracted and their corresponding source documents downloaded to form the relevant documents of the dataset. On the other hand, the irrelevant news articles consist of general health-related news, but without direct or indirect mentions of disease outbreaks (e.g., *plague*, *cholera*, *cough*), as well as general news like politics and sports. Most of the irrelevant documents were obtained from the News Category Dataset (Misra, 2018) comprising Huffpost[3] news articles for the period 2012 to 2018. The news articles cover various topics such as culture, politics, wellness, among other topics.

	All	Polish	Chinese	Russian	Greek	French	English
Train	5,074 (10.8)	241 (7.4)	300 (2.6)	296 (9.45)	253 (6.7)	1,593 (10.9)	2,365 (11.7)
Validation	1,250 (10.9)	54 (7.4)	71 (2.8)	60 (10.0)	68 (10.2)	388 (13.4)	583 (12.6)
Test	1,250 (10.5)	46 (13.0)	75 (6)	70 (10.0)	63 (4.7)	434 (12.4)	614 (12.8)

Table 2: The number of documents (percentage of relevant documents) per dataset split.

To simulate the real scenario of news reporting, we set the number of documents reporting disease outbreak (relevant documents) to be no more than 10% of the total dataset. The statistics of the dataset are presented in Table 1.

We split the data, with a total of $7,574$ articles, into training, validation, and testing sets. The training set comprises a total of $5,074$ documents, while the remaining documents were shared equally between the validation and the testing sets, that is, $1,250$ documents for validation, and $1,250$ documents for testing, stratified by language, as shown in Table 2. We also present the percent of relevant articles that refer to epidemiological news, which depicts the imbalanced nature of the dataset.

4 Experiments

The metrics considered in the evaluation of the models are precision, recall, and F1-score. Measuring recall is particularly important because of the risk posed by not identifying all the positive cases, with regard to disease outbreaks.

4.1 Models and Hyperparameters

Baseline model: DANIEL As a baseline model, we chose DANIEL[4] (Lejeune et al., 2015), an unsupervised system that does not rely on any language-specific grammar analysis and considers text as a sequence of strings instead of words. Consequently, DANIEL can be easily adapted to operate on any

[3] https://www.huffpost.com/

[4] https://github.com/NewsEye/event-detection/tree/master/event-detection-daniel

language and extract crucial information early on, which can significantly improve the decision-making process. This is pivotal in epidemic surveillance since timeliness is key, and more than often, initial medical reports are in the vernacular language where patient zero appears (Lejeune et al., 2015). We did not evaluate the BIOCASTER because only the ontology is publicly available and covers a limited number of languages, while GRITS is targeted to mostly English text.

Machine Learning models We also investigate three commonly used text classification models as baselines, Logistic Regression (LR), Random Forest (RF), and Support Vector Machine (SVM), using default hyperparameters and the TF-IDF weighting measure[5].

Deep Learning models Firstly, we consider two models, a CNN and a BiLSTM with FastText (Joulin et al., 2016) word representations for all the languages in the dataset, with an embedding dimension of 300. For the CNN, a sequence of word embeddings is passed through a convolution of kernel size 3 and a filter size of 250. Similarly, the BiLSTM passes the word embeddings through a bi-directional LSTM with a cell size of 128. Other hyperparameters for the models are a batch size of 32, a learning rate of 1×10^{-2}, and 15 epochs with early stopping of 3 to avoid overfitting.

Additionally, we chose to perform experiments with different BERT-based architectures (Devlin et al., 2018) for the sequence classification task. We used the default hyperparameters, a learning rate of 2×10^{-5}, and a maximum length of 512 tokens, with the longer sentences truncated to the defined maximum length. The pre-trained models are the `bert-base-multilingual-cased` and `uncased`. Finally, the CNN/BiLSTM described earlier in this section, but this time utilizing BERT features were also evaluated. We also test a graph convolutional networks (GCN) based-approach that augments BERT with graph embeddings (VGCN+BERT) (Lu and Nie, 2019). A GCN is a multilayer neural network that calculates directly on a graph and induces embedding vectors of nodes based on properties of their neighborhoods. Combining the capabilities of BERT with GCNs has been shown to be effective in capturing both local information and global information.

Models	**Precision %**	**Recall %**	**F1 %**
DANIEL	33.9	60.61	43.48
LR	93.81	68.94	79.48
RF	**95.70**	67.42	79.11
SVM	91.26	71.21	80
CNN+FastTtext	86.11	70.45	77.5
BiLSTM+FastTtext	77.44	78.03	77.74
BERT (`cased`)†	88.62	82.58	85.49
CNN+BERT (`cased`)†	88.79	71.97	79.5
BiLSTM+BERT (`cased`)†	90.20	69.70	78.63
BERT (`uncased`)†	84.67	87.88	**86.25**
CNN+BERT (`uncased`)†	82.14	87.12	84.56
BiLSTM+BERT (`uncased`)†	83.72	81.82	82.76
BERT (`cased`)	80.71	85.61	83.09
CNN+BERT (`cased`)	86.67	78.79	82.54
BiLSTM+BERT (`cased`)	75.95	**90.91**	82.76
BERT (`uncased`)	88.52	81.82	85.04
CNN+BERT (`uncased`)	86.07	79.55	82.68
BiLSTM+BERT (`uncased`)	81.51	73.48	77.29
VGCN+BERT	87.18	77.27	81.93

Table 3: Evaluation scores of the analyzed models for the relevant documents for all languages. The pre-trained BERT models are `base-multilingual`. LR stands for Logistic Regression, RF for Random Forest, and SVM for Support Vector Machines, †fine-tuned.

[5] `https://scikit-learn.org/`

4.2 Results

Deep learning transfer approaches such as BERT have demonstrated the ability to outperform the state-of-the-art methods on larger datasets. However, when there exist only a few labeled examples per class, 100 to $1,000$ as is the case of the low-resourced languages present in the dataset used in this study, the choice of the most appropriate approach is less clear, with classical machine learning and deep transfer learning presenting plausible options. Results of the experiments for different machine learning and deep learning models, using the dataset splits indicated in Table 2 are presented in Table 3 and discussed below.

Regarding the machine learning methods, we notice, from the results in Table 3, that SVM outperforms by a small margin the LR and RF on precision and recall, while the RF has not only the highest precision (95.70%) among this category of models, but the highest compared to all the models analyzed. We observe that the machine learning models (LR, RF, SVM) are greatly imbalanced, registering the highest values in precision and the lowest in the recall. This can be detrimental to the interests of an epidemiological detection system. Compared with the baseline results provided by DANIEL, this specialized model had a higher recall than precision which proves the specialized nature of such a tool, although its recall is the lowest among all the methods compared.

On the other hand, the models based on either CNN or BiLSTM with FastText embeddings have lower F1 scores than the classical machine learning methods (LR, RF, SVM). This could be explained by the fact that the training data is insufficient to train the models to have the ability to better distinguish between relevant and irrelevant documents.

In the case of deep transfer learning models, one can notice a great difference in the F1 score performance of BERT-based models, compared to all the other models. We can also observe that BERT-based models manage to balance recall and precision (precision remains consistent despite the increase in recall). The models benefit from the pre-trained language models that are either used as features or fine-tuned on the task. BERT relies on Byte Pair Encoding (BPE) based WordPiece tokenization (Wu et al., 2016) which makes it more robust to handle out-of-vocabulary words.

Models	Polish	Chinese	Russian	Greek	French	English
DANIEL	40	80	33.33	33.33	71.43	32.23
LR	0	0	66.67	66.67	84.21	80
RF	0	0	40	66.67	86.84	78.83
SVM	0	0	33.33	0	87.18	81.38
CNN+FastText	0	0	0	0	84.21	81.88
BiLSTM+FastText	0	0	0	0	73.12	85.71
BERT (`cased`)†	50	80	66.67	66.67	**94.12**	82.89
CNN+BERT (`cased`)†	50	80	66.67	40	86.05	86.75
BiLSTM+BERT (`cased`)†	0	80	40.00	66.67	87.36	86.27
BERT (`uncased`)†	57.14	80	50	**100**	91.95	86.08
CNN+BERT (`uncased`)†	50	80	66.67	40	86.05	86.75
BiLSTM+BERT (`uncased`)†	0	80	40	66.67	87.36	86.27
BERT (`cased`)	33.33	80	50	66.67	87.50	85.54
CNN+BERT (`cased`)	0	0	40	66.67	83.33	86.45
BiLSTM+BERT (`cased`)	0	80	22.22	28.57	85.11	**88.37**
BERT (`uncased`)	0	66.67	85.71	66.67	87.18	86.25
CNN+BERT (`uncased`)	0	50	40	66.67	82.35	86.45
BiLSTM+BERT (`uncased`)	0	0	33.33	0	72.94	84.42
VGCN+BERT	**71.43**	**88.89**	**88.89**	80	87.80	78.26

Table 4: F1-micro scores of the analyzed models for the relevant documents per language. The pre-trained BERT models are `base-multilingual`. LR stands for Logistic Regression, RF for Random Forest, and SVM for Support Vector Machines, †fine-tuned.

Regarding the difference between the fine-tuned BERT-based models and those that use the BERT encoder for generating features only, the performance is slightly better when BERT is fine-tuned on the task. However, in the case of additional layers on top of the BERT encoder, when fine-tuned, a considerable decrease in performance can be seen. Overall, these results suggest that the deep learning approaches are capable of much deeper and complex representations, such that they can utilize previously learned features for newer documents, even when the language of the document differs.

As observed in Table 4, all the machine learning models (LR, RF, SVM) display similar trends in their unequal performance based on language by not detecting (having the F1 values of zero) the relevant documents in Polish and Chinese. This is likely due to the size of the training data for these particular languages. Similarly, for all the low-resource languages (Polish, Chinese, Russian, and Greek), unsurprisingly, the CNN and BiLSTM -based models with pre-trained FastText embeddings were not able to distinguish relevant documents from irrelevant ones, as indicated by their low F1 scores. This might be due to the low embedding coverage of the languages. The F1 values for Chinese tend to be consistent for all BERT-based models while the performance for Polish varies a lot between models. VGCN+BERT had the highest F1 scores for the low-resourced languages Polish, Chinese, and Russian and the second-highest for Greek.

In order to analyze the influence of the documents with a larger quantity of documents (French and English, around $2,000$ news articles) over the classification of low-resource languages, we consider every language as the source language and the other five languages as target languages. At every iteration, the best performing model from the previous experiments is trained on the data in the source language and applied directly to every target language.

Test **Train**	**Polish**	**Chinese**	**Russian**	**Greek**	**French**	**English**
Polish	40	0	66.67	66.67	76.92	85.71
Chinese	0	80	60	0	70.97	81.08
Russian	33.33	0	33.33	66.67	62.86	88.61
Greek	0	0	0	66.67	0	63.05
French	0	66.67	57.14	0	91.95	85.90
English	50	0	33.33	66.67	39.29	84.35

Table 5: Evaluation scores of the BERT (`multilingual-uncased`)† fine-tuned model for the relevant documents in a zero-shot transfer learning setting.

The performance of models trained on the English and French documents is consistently higher than models trained on the other languages, as shown in Table 5. This can mainly be attributed to the larger quantity of annotated data ($> 2,000$ documents for training) for the two languages compared to the other languages. Also, English typology more closely resembles French typology as it has more recent influence from French and other Romance languages. The two languages share lexical similarities and cognate words. Looking at familial origins of the Slavic languages, Russian and Polish, the languages have typological properties that are intuitively more important for a model based on a language model. However, we noticed that their performance varies greatly in the case of Polish, and less in the case of Russian. Considering the quantity of training data, the difference of only around 50 more documents for Russian in train set compared with Polish seems to influence the performance.

4.2.1 Effect of Article Structure

In the approach presented by Lejeune et al. (2015), the document is considered as the main unit and it has language-independent organizational properties. The assumption is that the document-detectable features at a document granularity offer high robustness at the multilingual scale. The author suggests using the text as a minimal unit of analysis beyond its relation to the genre from which it came. The press article is thus of this type, which has precise rules: the structure of the press article and the vocabulary used are established and there are well-defined communication aims known to the source as well as the target of the documents. These rules, at a higher level than the grammatical rules, are very similar in different languages, and from the knowledge of these rules, remarkable positions are defined which are independent of languages. To exploit particular zones of news article content, we perform experiments similar to (Lejeune et al., 2015) inspired by the work on genre invariants carried out by Giguet and Lucas (2004) and Lucas (2009). The different areas of texts that we analyze are as follows:

- Beginning of the text: ideally composed of the title of the article
- Beginning of body: containing the first two paragraphs
- End of body (foot): comprising the last two paragraphs
- Rest of body: made up of the rest of the textual elements (e.g., paragraphs)

Text Position	**Models**	**Precision %**	**Recall %**	**F1 %**
Beginning	VGCN+BERT	**87.18**	77.27	81.93
	BERT (uncased)†	84.67	87.88	86.25
Body	VGCN+BERT	79.83	71.97	75.70
	BERT (uncased)†	75.71	80.30	77.94
End	VGCN+BERT	72.93	73.48	73.21
	BERT (uncased)†	76.12	77.27	76.69
Beginning+End	VGCN+BERT	86.61	83.33	84.94
	BERT (uncased)†	85.61	**90.15**	**87.82**

Table 6: Performance based on portions of the documents using the best performing model, BERT (uncased) fine-tuned and the VGCN-based model. The pre-trained BERT models are base-multilingual. All positions of text have a limit of 512 tokens.

The results, as presented in Table 6, indicate that the combination of the beginning and the concluding text in the news documents provided the best features required to classify a document as either relevant or irrelevant to a disease outbreak. The lowest performance score was noted when the body and the conclusion were evaluated independently.

4.2.2 Effect of Training Data Size

Different sizes of the training data were selected at an interval of ten percent and evaluated to ascertain the impact on the overall performance of the best model, in this case, the BERT (multilingual-uncased) fine-tuned model.

We observe that there is a generally positive trend for F1 score performance when trained on increasingly large datasets, as can be seen in Figure 1. When using only 10% of the data, the model achieves an F1 score performance that is comparable to that of the classical machine learning models and plateaus at 30% of the data. It is worth noting that the model achieves an F1 score of 64.03 using 5% of the training data, which is a significant performance for such a minimal amount of data.

5 Discussion

Out of all the models, deep learning BERT-based models were the best performing models, in terms of both F1 score and recall measures. The good performance can be attributed to the deep network architectures and large corpora used to train Transformer-based pre-trained language models (PMLs) such as BERT, which enable learning of rich text representations. Moreover, BERT fine-tuning performed better compared to the feature-based approaches, where FastText and BERT embeddings were used as input features to CNN and BiLSTM classifiers. Essentially, the PLMs end up learning universal language representations that are beneficial to downstream tasks.

The high precision and low recall noted in the machine learning models suggest that the models are unable to detect the relevant class well but are highly reliable when they do. This implies that while the

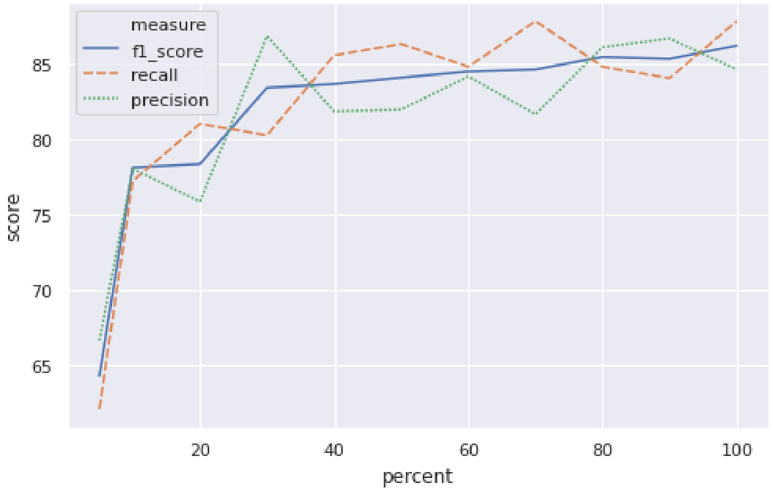

Figure 1: Impact of data size on performance of the best performing model: BERT (`multilingual-uncased`) fine-tuned.

classifiers returned reliable results, the machine learning models had a high false-negative rate, hence a few of all relevant results were returned. The approaches based on fine-tuned BERT uncased generally struck a good balance between precision and recall.

VGCN+BERT performed particularly well for Polish, Chinese, and Russian. The model utilizes graph embeddings produced by integrating local information captured by BERT and global information from the vocabulary graph that is based on word co-occurrence information. Both the local and global information interact with each other through a self-attention mechanism during the learning process. The interaction introduces useful global information to BERT, which contributes to the improved results across all the languages, including the low-resource languages.

With regard to the contribution of various document segments on performance, it was observed from the results that, the beginning and the end of the text combined had the highest recall and F1 score. This was particularly the case for models based on BERT namely, VGCN+BERT and BERT fine-tuned models. This can be explained by the fact that the beginning paragraphs in an article often capture the most important information, which informs the reader what the story is about. On the other hand, the last part of the article tends to provide a summary of the article.

The performance of the model improved proportionately with training data size. This is in line with neural network models, which require large amounts of data to train and evaluate. The competitive performance even with a small amount of data results from the transfer of knowledge from the pre-trained language model, trained on a large corpus, to the specific task of classifying epidemic text. This demonstrates the extent to which transfer learning can benefit the process of extracting useful information from multilingual epidemiological text.

6 Conclusions

Building effective epidemiological surveillance systems is of high importance these days. Detection of news reports on disease outbreaks is a crucial requirement of such systems. In this paper, we study in detail the performance of different methods on the task of epidemiological news report detection. The evidence presented in this work suggests that the models based on fine-tuned language models and/or graph convolutional networks achieve very good performance ($> 90\%$) on the classification of multilingual epidemiological texts, not only for high-resource languages but also for low-resource languages. In future work, we will consider the perspective of pursuing the task of epidemiological event extraction from news texts in low-resourced languages.

Acknowledgements

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

References

Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. 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.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Theresa Marie Bernardo, Andrijana Rajic, Ian Young, Katie Robiadek, Mai T Pham, and Julie A Funk. 2013. Scoping review on search queries and social media for disease surveillance: a chronology of innovation. *Journal of medical Internet research*, 15(7):e147.

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

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

Lauren E Charles-Smith, Tera L Reynolds, Mark A Cameron, Mike Conway, Eric HY Lau, Jennifer M Olsen, Julie A Pavlin, Mika Shigematsu, Laura C Streichert, Katie J Suda, et al. 2015. Using social media for actionable disease surveillance and outbreak management: a systematic literature review. *PloS one*, 10(10):e0139701.

Nigel Collier, Son Doan, Ai Kawazoe, Reiko Matsuda Goodwin, Mike Conway, Yoshio Tateno, Quoc-Hung Ngo, Dinh Dien, Asanee Kawtrakul, Koichi Takeuchi, et al. 2008. Biocaster: detecting public health rumors with a web-based text mining system. *Bioinformatics*, 24(24):2940–2941.

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

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

Crystale Purvis Cooper, Kenneth P Mallon, Steven Leadbetter, Lori A Pollack, and Lucy A Peipins. 2005. Cancer internet search activity on a major search engine, united states 2001-2003. *Journal of medical Internet research*, 7(3):e36.

Aron Culotta. 2010. Detecting influenza outbreaks by analyzing twitter messages. *arXiv preprint arXiv:1007.4748*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Ernesto Diaz-Aviles, Avaré Stewart, Edward Velasco, Kerstin Denecke, and Wolfgang Nejdl. 2012. Epidemic intelligence for the crowd, by the crowd. In *Sixth International AAAI Conference on Weblogs and Social Media*.

Son Doan, Quoc-Hung Ngo, Ai Kawazoe, and Nigel Collier. 2019. Global health monitor: A web-based system for detecting and mapping infectious diseases. *arXiv preprint arXiv:1911.09735*.

Emmanuel Giguet and Nadine Lucas. 2004. La détection automatique des citations et des locuteurs dans les textes informatifs. *Le discours rapporté dans tous ses états: Question de frontières*, pages 410–418.

Jeremy Ginsberg, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski, and Larry Brilliant. 2009. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2015. Bilbowa: Fast bilingual distributed representations without word alignments.

Felix Hamborg, Soeren Lachnit, Moritz Schubotz, Thomas Hepp, and Bela Gipp. 2018. Giveme5w: Main event retrieval from news articles by extraction of the five journalistic w questions. 03.

Andrew G Huff, Nathan Breit, Toph Allen, Karissa Whiting, and Christopher Kiley. 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.

Aditya Joshi, Sarvnaz Karimi, Ross Sparks, Cécile Paris, and C Raina Macintyre. 2019. Survey of text-based epidemic intelligence: A computational linguistics perspective. *ACM Computing Surveys (CSUR)*, 52(6):1–19.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651.*

Dan Kondratyuk and Milan Straka. 2019. 75 languages, 1 model: Parsing universal dependencies universally. *arXiv preprint arXiv:1904.02099.*

Sneha Kudugunta, Ankur Bapna, Isaac Caswell, and Orhan Firat. 2019. Investigating multilingual NMT representations at scale. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1565–1575, Hong Kong, China, November. Association for Computational Linguistics.

Alex Lamb, Michael J Paul, and Mark Dredze. 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.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291.*

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

Jiwei Li and Claire Cardie. 2013. Early stage influenza detection from twitter. *arXiv preprint arXiv:1309.7340.*

Zhibin Lu and Jian-Yun Nie. 2019. Raligraph at hasoc 2019: Vgcn-bert: Augmenting bert with graph embedding for offensive language detection.

Nadine Lucas. 2009. *Modélisation différentielle du texte, de la linguistique aux algorithmes*. Ph.D. thesis, Université de Caen.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 151–159.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168.*

Rishabh Misra. 2018. News category dataset. https://www.kaggle.com/rmisra/news-category-dataset.

Stephen Mutuvi, Antoine Doucet, Gaël Lejeune, and Moses Odeo. 2020. A dataset for multi-lingual epidemiological event extraction. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4139–4144, Marseille, France, May. European Language Resources Association.

Michael J Paul, Abeed Sarker, John S Brownstein, Azadeh Nikfarjam, Matthew Scotch, Karen L Smith, and Graciela Gonzalez. 2016. Social media mining for public health monitoring and surveillance. In *Biocomputing 2016: Proceedings of the Pacific symposium*, pages 468–479. World Scientific.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502.*

Philip M Polgreen, Yiling Chen, David M Pennock, Forrest D Nelson, and Robert A Weinstein. 2008. Using internet searches for influenza surveillance. *Clinical infectious diseases*, 47(11):1443–1448.

Libo Qin, Minheng Ni, Yue Zhang, and Wanxiang Che. 2020. Cosda-ml: Multi-lingual code-switching data augmentation for zero-shot cross-lingual nlp. *arXiv preprint arXiv:2006.06402.*

Adam Sadilek, Henry Kautz, and Vincent Silenzio. 2012. Predicting disease transmission from geo-tagged microblog data. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*.

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

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.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. *arXiv preprint arXiv:1904.09077*.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.