Information Processing and Management 56 (2019) 247–257

Contents lists available at ScienceDirect

Information Processing and Management

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

A tale of two epidemics: Contextual Word2Vec for classifying twitter streams during outbreaks

Aparup Khatua[a,b], Apalak Khatua[c], Erik Cambria[b,*]

[a] Department of Computer Science and Engineering, University of Calcutta, India
[b] School of Computer Science and Engineering, Nanyang Technological University, Singapore
[c] XLRI – Xavier School of Management, Jamshedpur, India

ARTICLE INFO

Keywords:
Epidemics
Ebola
Zika
PubMed
Twitter
Text classification
Word Vectors

ABSTRACT

Unstructured tweet feeds are becoming the source of real-time information for various events. However, extracting actionable information in real-time from this unstructured text data is a challenging task. Hence, researchers are employing word embedding approach to classify unstructured text data. We set our study in the contexts of the 2014 Ebola and 2016 Zika outbreaks and probed the accuracy of domain-specific word vectors for identifying crisis-related actionable tweets. Our findings suggest that relatively smaller domain-specific input corpora from the Twitter corpus are better in extracting meaningful semantic relationship than generic pre-trained Word2Vec (contrived from Google News) or GloVe (of Stanford NLP group). However, domain-specific quality tweet corpora during the early stages of outbreaks are normally scant, and identifying actionable tweets during early stages is crucial to stemming the proliferation of an outbreak. To overcome this challenge, we consider scholarly abstracts, related to Ebola and Zika virus, from PubMed and probe the efficiency of cross-domain resource utilization for word vector generation. Our findings demonstrate that the relevance of PubMed abstracts for the training purpose when Twitter data (as input corpus) would be scant during the early stages of the outbreak. Thus, this approach can be implemented to handle future outbreaks in real time. We also explore the accuracy of our word vectors for various model architectures and hyper-parameter settings. We observe that Skip-gram accuracies are better than CBOW, and higher dimensions yield better accuracy.

1. Introduction

Twitter, a microblogging platform, allows its user to post *tweets,* which are short text messages up to 140 characters in length. A user can interact with others by replying or retweeting. Over the years, the Twitter platform is becoming a real-time communication channel during various natural disasters such as tsunamis, hurricanes, earthquakes, floods, and epidemics. For instance, the 2010 Haiti earthquake survivors communicated among themselves through social networks, such as Twitter and Facebook, as an alternate option of the jammed mobile phone network. More importantly, communication via these social media platforms raised a donation of US$8 million within 48 hours of the Haiti earthquake (Gao, Barbier, & Goolsby, 2011). A sudden outbreak of an epidemic can lead to a crisis where the administration has to take quick decisions in the absence of structured data. Thus, the challenge is to develop a framework for the effective information retrieval, from unstructured social media data in real-time, to augment the traditional relief

* Corresponding author.
E-mail addresses: akhatua@ntu.edu.sg (A. Khatua), apalak@xlri.ac.in (A. Khatua), cambria@ntu.edu.sg (E. Cambria).

https://doi.org/10.1016/j.ipm.2018.10.010
Received 5 November 2017; Received in revised form 29 August 2018; Accepted 12 October 2018

operations (Imran, Elbassuoni, Castillo, Diaz, & Meier, 2013). This requires systematic collection and conversion of the unstructured data into structured data by theme-wise classification and aggregation of tweets.

Nowadays researchers are employing word embedding for information retrieval from unstructured text corpora (Choi et al., 2016; Fernández-Reyes, Hermosillo-Valadez, & Montes-y-Gómez, 2018; Glorot, Bordes, & Bengio, 2011; Poria et al., 2012). Similarly, in the context of the 2014 Ebola and 2016 Zika outbreaks, we employ word embedding for classification of Twitter corpora. This vector space representation of words allows researchers to probe the semantic and syntactic pattern within an unstructured text corpus (Mikolov, Chen, Corrado, & Dean, 2013a; 2013b). In the biomedical domain a very few studies employed word vectors contrived from electronic health records for information retrieval (Choi et al., 2016; Cambria et al., 2010a). Choi et al. (2016) argued that this domain-specific approach is efficient not only for document classification but also for medical diagnosis and medication. Therefore, this paper investigates whether disease-specific word vectors outperforms pre-trained generic word vectors such as *Word2Vec* (contrived from Google news) or Global Vectors for Word Representation (commonly known as *GloVe*) constructed by Pennington, Socher, and Manning (2014).

Prior studies, such as Ghosh, Chakraborty, Cohn, Brownstein, and Ramakrishnan (2016), considered Twitter corpus for preparing disease-specific word embedding, but constructing efficient word vector model requires a voluminous domain rich text corpus for the training purpose (Chiu, Crichton, Korhonen, & Pyysalo, 2016). Prevalence of misinformation and propagation of rumors on Twitter platforms compromise the quality of Twitter data (Ashktorab, Brown, Nandi, & Culotta, 2014; Jin et al., 2014; Kalyanam, Velupillai, Doan, Conway, & Lanckriet, 2015). Hence, Twitter data requires time-consuming manual pre-processing. Moreover, Twitter discussion gains momentum only after the crisis reaches a threshold point. Thus, generating a voluminous domain rich text corpus for the training purpose is problematic during the initial stages of the crisis. Prior studies in the biomedical domain considered scholarly abstract for creating word vectors (Turner et al., 2017; Zhu et al., 2017). Thus, in addition to tweet corpus, we have also considered scholarly abstracts from Ebola and Zika virus-related articles from the biomedical literature for constructing word vectors.

Our findings suggest that scholarly abstracts are a viable alternative for creating the word embedding. First, unlike tweet feeds, scholarly abstracts are not fraught with the pitfall of misinformation or rumors. Second, throughout the history of humanity epidemics are responsible for thousands of lives lost. For instance, Ebola first emerged in Sudan and Zaire back in 1976 (Source: WHO). Thus, a plethora of biomedical literature on Ebola was available before the 2014 Ebola crisis. Hence, this approach can be extrapolated for other outbreaks in future. To sum up, unlike tweet feeds, availability of scholarly abstracts (both quality and quantity wise) does not depend on the severity and the stages of the outbreak.

For the sake of robustness, we have considered various context window sizes (W) as well as different dimension sizes (D) for our analysis. We observe that disease-specific word embedding is more accurate than other pre-trained generic word vectors. Our results broadly remain consistent, across different W and D, for both the outbreaks. Thus, this study elucidates that scholarly abstracts from biomedical literature can be an alternate cost-effective and time-independent option for input corpora. Establishing the relevance of scholarly abstracts for classification of the unstructured Twitter corpus, during infectious disease outbreaks, is the core contribution of the study.

The remainder of the paper is organized as follows: Section 2 discusses the prior work on information retrieval using Twitter data and posits our research gap. Section 3 elaborates our research context and highlights the idiosyncratic nature of two epidemics considered in this study. We describe the experimental design and methodology in Section 4. In Section 5, our main results are presented and discussed. We also highlight the implications of our work in this section. Section 6 concludes the paper by chalking out the future scope of work.

2. Information retrieval using twitter data during outbreaks

World Health Organization (WHO) defines syndromic surveillance as 'the continuous, systematic collection, analysis, and interpretation of health-related data needed for the planning, implementation, and evaluation of public health practice.' Nowadays, web-based contents, ranging from social media platforms to web-based search query data, are generating an enormous amount of user-generated data. Prior studies have investigated Google (Chan, Sahai, Conrad, & Brownstein, 2011; Ginsberg et al., 2009), Wikipedia (Hickmann et al., 2015), Twitter (Culotta, 2010; Oyeyemi, Gabarron, & Wynn, 2014) to monitor and forecast potential health concerns such as influenza-like illness (Ghosh et al., 2016), Ebola outbreak (Chan et al., 2011; Ginsberg et al., 2009), Zika Virus (Khatua & Khatua, 2016; Miller, Banerjee, Muppalla, Romine, & Sheth, 2017) and so on. Traditional syndromic surveillance efforts can be augmented by efficient information extraction from various web-based contents. Google Flu Trend was one of the initial attempts to use web search query data for detecting outbreaks. Some scholars argue that web search query volume can predict the emergence and spread of infectious diseases such as influenza or dengue (Chan et al., 2011; Ginsberg et al., 2009). However, Lazer, Kennedy, King, and Vespignani (2014) argue that Google Flu Trend significantly overestimate the prevalence of flu and labeled it as *big data hubris*. It is worth noting that Google trend only reveals the volume and location of the query, and it lacks the richness of microblogging platforms such as Twitter.

Twitter has emerged as one of the popular platforms for users not only to communicate with each other but also to share information with others. Millions of netizens 'are searching online health information and also publishing content about their health' (Fernández-Luque & Bau, 2015). Consequently, social media platforms became a real-time source for information retrieval during an outbreak. Extant literature explored Twitter data during global health crises such as the 2009 H1N1 pandemic (Chew & Eysenbach, 2010; Culotta, 2010), the 2014 Ebola outbreak (Jin et al., 2014; Kalyanam et al., 2015; Oyeyemi et al., 2014; Rudra, Sharma, Ganguly, & Imran, 2018), and the 2016 Zika outbreak (Khatua & Khatua, 2016; Miller et al., 2017). Initial Twitter-based works mostly performed volumetric analysis to predict the propagation of an outbreak. For instance, prior studies noted a strong correlation

between influenza-related tweet volume and patient statistics data of the Centers for Disease Control and Prevention (CDC) (Culotta, 2010). However, this approach does not consider the text content of tweet feeds.

It is worth noting that Twitter data is unstructured. More importantly, microblogging platforms are 'often overwhelmed by irrelevant and redundant noise' (Ashktorab et al., 2014) and can propagate medical misinformation (Jin et al., 2014; Kalyanam et al., 2015). Therefore, the challenge is to develop an efficient information retrieval framework for extracting useful and actionable tweets during a crisis event to know what is happening on the ground to augment the relief operation (Vieweg et al., 2010; Imran et al., 2013). Hence, Fernández-Luque and Bau (2015) suggest developing analytical tools for investigating these unstructured data which will be beneficial for healthcare professionals and regulatory authorities.

Natural language processing techniques are capable of clustering, named entity recognition, automatic summarization, information classification (supervised and unsupervised) and information retrieval tasks (Cambria, Mazzocco, Hussain, & Eckl, 2011). Employing word embedding can enhance the efficiency of meaningful information retrieval. For instance, Word2Vec models, proposed by Mikolov et al. (2013a; 2013b), consider a corpus as input and create a vector space. Word2Vec represents each unique word as a vector, and contextual similarity of two words can be inferred from these vectors (Zhu et al., 2017; Cambria, Hussain, Durrani, Havasi, & Eckl, 2010b). In other words, the cosine similarity of two vectors (representing two words) indicates the semantic relationship between two words. Classical text classification process fails to consider the semantic relationship between words. Understanding semantic relationships are important for identifying latent themes within the Twitter corpus during the crisis.

Only a handful of studies, in the biomedical context, have employed word embedding to investigate similarity and semantic relatedness for information retrieval and extraction (such as Minarro-Giménez, Marin-Alonso, & Samwald, 2014; TH, Sahu, & Anand, 2015; Turner et al., 2017; Zhu et al., 2017). However, a few of them have considered PubMed abstracts as input corpora (Chiu et al., 2016; Minarro-Giménez et al., 2014; Pakhomov et al., 2010; TH et al., 2015). Couple of other studies considered domain-specific Word2Vec (by using disease-specific Twitter data as input corpora) for classification of tweet feeds (Ghosh et al., 2016; Rudra et al., 2018). However, none of the prior studies, to the best of our knowledge, has employed PubMed-based word vectors for the classification of tweet feeds during outbreaks.

2.1. Research gaps

On the basis of extant literature, it can be argued that word embedding on twitter data will allow a better understanding of underlying latent themes. Subsequently, word vectors will also enhance the information retrieval process from the surveillance perspective (Lampos, Zou, & Cox, 2017). Generic word vectors, constructed from the voluminous corpus, are not domain specific and can have broad application. More importantly, Chiu et al. (2016) pointed out that 'bigger corpora do not necessarily produce better biomedical domain word embeddings.' Ghosh et al. (2016) also echoed this view and argued that domain specific word vector models, instead of generic Word2Vec models, are more efficient to extract disease-specific information. Motivated by these arguments, we investigate whether a domain specific word vector model would be more appropriate for information retrieval during outbreaks. Thus, our first research question (RQ) is,

RQ 1. *: Whether constructing a Word2Vec from domain-specific Twitter corpus outperforms other pre-trained generic word vectors?*

However, word vectors require not only quantity but also quality text as input corpora (Chiu et al., 2016), and Twitter deliberation gains momentum only when the crisis reaches its peak. Thus, employing Twitter as input corpora might lead to a lag effect in relief operations. Additionally, misinformation and propagation of rumors on social media platforms have its shortcomings (Ashktorab et al., 2014; Jin et al., 2014; Kalyanam et al., 2015). Therefore, considering Twitter corpus for constructing word vectors in real-time for relief operations is problematic. To overcome this, Rudra et al. (2018) explored cross-domain analysis and investigated whether 'the classifier trained over tweets posted during a past outbreak' can 'predict tweets of a future/current outbreak.' This cross-domain approach might be applicable for same disease or diseases with similar symptoms and effects (such as dengue and influenza). However, epidemics are not homogenous. For instance, Ebola outbreaks are generally associated with a high mortality rate. On the contrary, the mortality rate for Zika is negligible, but it can have a disastrous effect on pregnant women and their babies (Miller et al., 2017). Twitter discussions regarding congenital disabilities or neurological disorders are specific to Zika outbreaks (Khatua & Khatua, 2016). Therefore, word vector models from Ebola corpus will be inefficient for understanding the semantic and syntactic relation of Zika corpus. This leads to our follow-up research question,

RQ 2. *: What can be the alternate source of text corpus for preparing the word vectors during the initial stages of the outbreak?*

To answer this, following the extant literature we explore PubMed, a publicly available source, which 'comprises more than 28 million citations for biomedical literature from MEDLINE, life science journals, and online books' (Source: PubMed website). We have extracted Ebola and Zika-related scholarly abstracts, from PubMed, to build our word vector models and compare its efficiency with other word vectors.

3. Research context: 2014 Ebola and 2016 Zika outbreaks

We set our study in the context of the 2014 Ebola and 2016 Zika outbreaks. On Aug 08, 2014, WHO declared Ebola as an international health emergency, and the death toll exceeded 1000 within a week. To tackle the situation, the WHO team went to these African countries with medical equipments and hygiene kits. This outbreak initially gathered momentum in Guinea, and subsequently, it spread to neighboring countries like Liberia and Sierra Leone. The United States, Australia, and many others countries also joined WHO in the relief operations. According to various media sources, the reported cases in five West African countries became

4293 with at least 2296 deaths by September 6, 2014. The situation rapidly became more and more severe, and on October 1, 2014, WHO confirmed 3338 deaths out of 7178 reported cases mostly in Guinea, Sierra Leone, and Liberia. Unfortunately, a significant portion of all reported cases ended in death in this deadliest outbreak. CDC was forced to suggest quarantine for clinicians returning from West Africa. In the middle of November 2014, WHO announced 5160 deaths across eight countries. The situation started to become normal only in early 2015.

Our second research context was the 2016 Zika outbreak. In November 2015, Brazil government declared Zika as a public health emergency. Two months later, WHO also declared Zika a public health emergency of international concern. In February 2016, the CDC announced that the Zika virus had been transmitted to 30 countries. On February 12, 2016, there were 462 confirmed cases of microcephaly, and 41 of them were linked to the Zika virus. Subsequently, in March 2016, the WHO informed that Zika could cause microcephaly (a congenital disability) for newborn babies as well as Guillain-Barré syndrome (a neurological disorder) for adults. Therefore, WHO suggested pregnant women to avoid traveling to Zika-affected areas, and also warned about the possibility of sexual transmission of Zika Virus (Khatua & Khatua, 2016; Miller et al., 2017).

It is worth noting that the health risks associated with Ebola and Zika outbreaks were different. Mostly epidemics, even the Ebola outbreaks, are associated with high death tolls. For example, the 2009 H1N1 virus infection 'resulted in an estimated range of deaths from between 151,700 and 575,400 people who perished worldwide … during the first year the virus circulated' (Source: CDC Influenza Division). However, the Zika virus disease mostly resembles mild fever, and generic drugs can be recommended for the treatment of infected patients. In the short run, unlike the Ebola or H1N1 pandemics, the mortality rate for Zika virus would be negligible. However, in the long run, the Zika virus can lead to an unprecedented rise of microcephaly and Guillain-Barré syndrome (Source: WHO).

One potential shortcoming of choosing developing countries as a research context is their low internet penetration rate. For instance, the internet penetration was around mere 10% in Liberia or Nigeria during the outbreak. On the contrary, the internet penetration in developed economies, such as the United States or the United Kingdom, is over 85%. Internet penetration rate of around 55% in Brazil was significantly better than its African peers, but not at par with the developed economies. More importantly, the US has over 300 million citizens while Liberia has just over 4 million. Consequently, the volume and richness of user-generated Twitter data for influenza-like illness in the United States will be much better than the Ebola outbreak in African countries. However, in recent times the growth of Twitter has reached a stagnation point in the US and Europe. On the contrary, the number of active social media users, as well as internet penetration rate, have recorded phenomenal rise in developing countries. Twitter's growth has also shifted to developing economies from developed economies. Therefore, our proposed approach will become more and more relevant in developing economies with each passing day.

4. Methodology

4.1. Data collection

Twitter data: Our Ebola data comprises 0.60 million unique tweets, containing the keyword *Ebola*, posted between August 22, 2014, and October 21, 2014. In our pre-processing stage, we have removed URL and non-ASCII characters. We have also removed numbers, punctuations, and mentions from the tweets. Our Ebola corpus boils down to 0.25 million tweets after removing retweets. Finally, we have considered 0.095 million tweets with 15 or more words for constructing the word vector model. Similarly, our Zika data comprises 0.97 million unique tweets, containing the keywords *Zika* or *ZikaVirus*, posted between January 25, 2016, and February 14, 2016. After pre-processing, we ended up with 0.35 million tweets. Finally, we have considered 0.127 million tweets with more than 15 words for constructing the word vector model.

Scholarly abstracts from PubMed: PubMed is a publicly available NCBI resource of the biomedical literature. We have considered 4200 and 1919 Ebola-related and Zika-related abstracts respectively from this database for constructing our word vector models.

4.2. Classification and preparation of gold standard

We have referred CDC and WHO websites to identify various possible categories of Twitter discussions associated with 2014 Ebola and 2016 Zika outbreaks. In accordance with the CDC website, we have identified 5 different categories as follows: health risks, prevention, symptoms, transmission, and treatment (see Table 1). Twitter deliberations related to chances of death (in the context of Ebola) or congenital disabilities/neurological disorders (in the context of Zika) are depicted as *Health Risks* related tweets. Microblogging conversations about the prevention and precaution of Ebola and Zika virus/outbreaks are categorized as *Prevention* related tweets. Similarly, Twitter discussion/queries about common Ebola/Zika signs such as fever, rashes, and headaches, are classified as *Symptom* related tweets. Tweets are classified as *Treatment*-related tweets if the deliberations are about the results of treatment, the absence of a specific medicine/treatment or the need for research to develop a vaccine. Finally, microblogging discussions are categorized as *Transmission* related tweets if they are deliberating about modes of diffusion through mosquito bites (for Zika) or by direct contact with a person who is sick or from the mortal remains of a person who has died from Ebola. This categorization is in sync with prior studies such as Miller et al. (2017) and Rudra et al. (2018). A few sample tweets of each category is reported in Table 2. Following prior studies (Ghosh et al., 2016; Miller et al., 2017), we have not considered unrelated/generic tweet category either in our annotation or our analysis.

We have 518 and 840 annotated tweets for Ebola and Zika respectively (in 5 different categories as mentioned above) to train the classifier. This gold standard is comprised of tweets where annotators have unanimously agreed about the classification. Table 3

Table 1
Classification of tweets.
Source: CDC website

Labels	2014 Ebola outbreak	2016 Zika outbreak
Health risks	-	-
	Very high percentage of deaths among reported cases	Congenial disabilities such as microcephaly
		-
		Neurological disorders such as Guillain-Barré Syndrome
Prevention	-	-
	Practicing careful hygiene	Avoiding mosquito bites
	-	-
	Avoiding contact with an infected person's blood or body fluids	Taking precaution during sex
	-	-
	Avoid contact with bats and nonhuman primates	Avoiding traveling to Zika affected areas
Symptoms	-	-
	Symptoms, such as fever, headache, muscle and stomach pain, diarrhoea, vomiting, haemorrhage, which may appear after exposure to Ebola Virus	Fever, rash, headache, joint pain, conjunctivitis (red eyes), muscle pains and so on
Transmission	-	-
	Through direct contact with a person who is sick with or has died from Ebola	Through mosquito bites, blood transfusion, and sex
	-	-
	Through contaminated needles and syringes	From infected pregnant mother to child
	-	
	Through infected fruit, bats or primates	
Treatment	-	-
	No FDA approved vaccine was available for Ebola	No specific medicine for Zika
	-	-
	Thus, basic interventions such as providing intravenous fluids and balancing body salts	Thus, basic treatments such as medicines to reduce fever and pain, fluids to prevent dehydration

Table 2
Sample tweets from gold standards for various classes.

Class	Event	Sample tweets
Health risks	Ebola	*The numbers you have heard for the scale of Ebola cases deaths in Africa grossly underestimate the reality sources tell me today*
	Zika	*A virus doesn't have to be fatal to cause serious concern I may not die from Zika but I don't want my baby to have microcephaly*
Prevention	Ebola	*I hope yall are still washing your hands and practicing Ebola prevention if it can reach the US it can visit Nigeria again*
	Zika	*Zikavirus is not Ebola it needs a vehicle to spread we know this vehicle is the Aedes aegypti mosquito we can control this vector*
Symptoms	Ebola	*The Ebola virus takes days before you actually start having the symptoms so you may get tested for it before it actually strikes*
	Zika	*After reading a list of symptoms mild fever headache rash joint pain I have come to the conclusion that I may have Zika Zikavirus*
Transmission	Ebola	*Ebola virus can only be transmitted through an Ebola patient that has manifested the symptoms good hygiene can help control this EVD*
	Zika	*Hard to say how many Zika cases have occurred due to sexual transmission so far but I bet it's tiny compared to vector transmission*
Treatment	Ebola	*The full recovery to date of one infected contact in Nigeria is good news Ebola is not invariably a death sentence*
	Zika	*I' m not much a vaccine person only get compulsory ones yellow fever and am on a Vitamine B treatment to prevent Zika*

Table 3
Number of tweets in our gold standard.

Event	Prevention	Transmission	Symptoms	Health risks	Treatment	Total
Ebola	116	135	89	80	98	518
Zika	228	179	126	244	63	840

reports the number of tweets in our gold standard. Tweets in treatment categories are scant with respect to other categories. Miller et al. (2017) also noted that 'there were hardly any tweets about treatment, which was not surprising because there is currently no treatment for Zika.'

4.3. Data analysis

Machine learning algorithms, for text mining, cannot process raw text in string format. This limitation can be addressed by converting raw texts into numerical representations to perform various text processing such as classification, regression analysis, information extraction, sentiment analysis (Cambria, Poria, Gelbukh, & Thelwall, 2017). In the *word embedding* method, a word is represented by vectors of real numbers, commonly known as vector space representation. This vector space representation of words enables data analysis tasks such as information retrieval, named entity recognition, and document classification. Mostly word embedding methods outperform traditional approaches. Mikolov et al. (2013a)) showed that *Word2Vec* displays interesting semantic

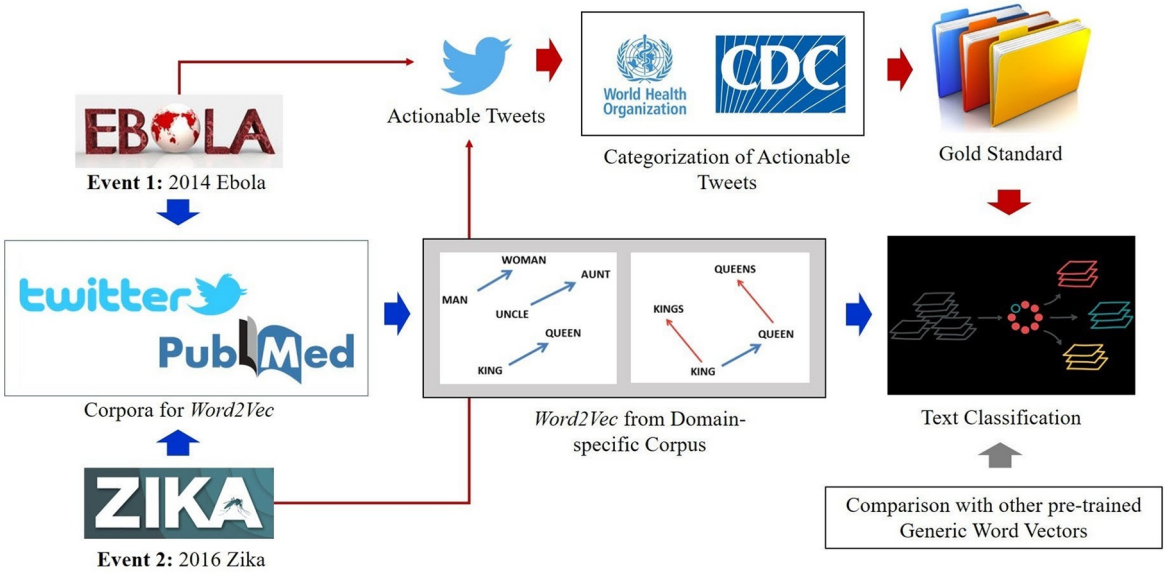

Fig. 1. Our research framework.

and syntactic relationships. For instance, the analogy *"king is to queen as man is to woman"* is encoded in a vector space by the vector equation as follows: *king − queen = man – woman*. This approach can have various practical applications. For instance, Fernández-Reyes et al. (2018) used word embedding for query expansion methods, and their novel word embedding approach has improved the overall performance of the global query expansion strategy. In brief, this vector space representation approach is 'highly efficient in finding word embedding templates from huge amounts of unstructured text data and uncover various semantic and syntactic relationships' (Ghosh et al., 2016).

Pennington et al. (2014) identified two basic types of learning word vectors as follows: traditional global matrix factorization methods, such as latent semantic analysis, and the local context window methods of Mikolov, Yih, and Zweig (2013b). However, Pennington et al. (2014) also pointed out that global matrix factorization methods 'efficiently leverage statistical information' but poor in word analogy task. On the contrary, Skip-gram model of Mikolov et al. (2013b) is efficient for word analogy task but fails to incorporate the statistical information. Thus, Pennington et al. (2014) proposed 'a specific weighted least squares model that trains on global word-word co-occurrence counts and thus makes efficient use of statistics.' This is commonly known as *GloVe*, and these pre-trained word vectors (of different dimensions) are publicly available. However, it is worth noting that extant literature has rarely considered the application of *GloVe* in the biomedical context (Zhu et al., 2017). For the sake of robustness, in our analysis we consider both pre-trained *Word2Vec* of Mikolov et al. (2013a; 2013b) as well as pre-trained *GloVe* of Pennington et al. (2014); and compare these models with our disease-specific *word vectors* (contrived from the Twitter corpus and PubMed scholarly abstracts). In other words, we are comparing our contextual *word vectors* (constructed from disease-specific text corpus) with freely available pre-trained word vectors, namely *Word2Vec* and *GloVe*.

Fig. 1 illustrates our overall research design. On the basis of our Twitter corpus and abstracts of biomedical literature from PubMed, we have generated 4 sets of *word vectors* as follows:

1 *Word2Vec* from Ebola-related tweet corpus
2 *Word2Vec* from Ebola-related abstracts from biomedical literature
3 *Word2Vec* from Zika-related tweet corpus
4 *Word2Vec* from Zika-related abstracts from biomedical literature

Prior studies noted that long texts are more informative than short texts. Thus, we have considered tweets which have minimum 15 words for preparing the word vectors from the Twitter corpora. However, this long vis-à-vis short text was not a constraint for preparing the word vectors from relatively longer PubMed abstracts.

Furthermore, we have considered two variations of *Word2Vec* as follows: Continuous Bag of Words (CBOW) model and Skip-gram model. CBOW predicts the target word on the basis of its contexts (or surrounding words). Thus, in CBOW the position of a particular word in the neighboring words does not matter. On the contrary, the Skip-gram model predicts surrounding words/context in a sentence for an input word (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013c). However, predicting context 'for all words for every training instance is very expensive, making it impractical to scale up to large vocabularies or large training corpora' (Rong, 2014). Thus, to address this problem Mikolov et al. (2013c) suggested *hierarchical softmax algorithm* which 'uses a binary tree representation of the output layer with the W words as its leaves and, for each node, explicitly represents the relative probabilities of its child nodes.' Hence, we also employed a hierarchical softmax algorithm for our analysis.

We use the *Word2Vec*[1] implementation in *Gensim* for generating a set of word vectors. Chiu et al. (2016) argue that 'the quality of word embeddings' depends not only on input corpora but also on hyper-parameter settings. Thus, we have considered various window size W and the dimensionality D (Yang, Macdonald, & Ounis, 2016). Following prior studies, we have considered three context window size as follows: W = {3, 5, 8} for CBOW and W = {5, 10, 15} for Skip-gram. Let's say a sentence comprises n-words as follows: w_1, w_2, … w_n. Thus, the context of a word w_i in this sentence comes from a window of size k around the word: $C(w) = w_{i-k}, \ldots, w_{i-1}, w_{i+1}, \ldots, w_{i+k}$, where k is a parameter (Goldberg & Levy, 2014). Furthermore, for each W we consider three different dimension size as follows: D = {50, 100, 200} to cover the low to high dimensionalities of word vectors. The dimensionality of a word vector indicates the dimension (D) of vector size. For instance, D = {200} indicates that for every word the vector will consider the relationship with another 200 corresponding words. Mikolov et al. (2013a), in their seminal work, argue that prior studies mostly considered 'a few hundreds of millions of words, with a modest dimensionality of the word vectors between 50 – 100.' However, they noted that when they 'train high dimensional word vectors on a large amount of data, the resulting vectors can be used to answer very subtle semantic relationships between words… Word vectors with such semantic relationships could be used to improve many existing NLP applications, such as … information retrieval.' However, Mikolov et al. (2013a) also noted that beyond an optimal point higher D might not lead to higher performance because the dimension (D) of a word vector depends on factors such as the total number of the unique words present in the training text, the length of each tuple used for classification, and so on. Thus, context-specific input corpora with low D can outperform a generic word2vec with high D.

It is worth noting that word vectors can be of two types namely, count-based and TF-IDF-based (term frequency-inverse document frequency). TF-IDF assigns higher importance to less occurred but important/crucial words which are potentially more meaningful to understand the context. Hence, in accordance to prior studies, we have considered both count-based and TF-IDF based word vectors in our analysis (Lilleberg, Zhu, & Zhang, 2015; Wang et al., 2015). To sum up, we are generating multiple types of word vectors by using different combinations of W and D, and finally comparing these models with publicly available pre-trained generic word vectors.

We have considered the *extra tree classifier* for our classification task. An extra-tree classifier 'implements a meta estimator that fits a number of randomized decision trees (a.k.a. extra-trees) on various sub-samples of the dataset and use averaging to improve the predictive accuracy and control over-fitting' (Source: http://scikit-learn.org). In other words, these ensemble methods are superior to a single classifier because these are 'learning algorithms that construct a set of classifiers and then classify new data points by taking a (weighted) vote of their predictions' (Dietterich, 2000). We have considered the value of *n_estimator* (i.e., the total number of trees for estimation) as 200 for our analysis. In addition to this *n_estimator*, we have considered default parameters for all other input variables (refer: http://scikit-learn.org for details of other default parameters). We have trained our model by a relatively smaller set of manually annotated data, and the lack of voluminous data is a constraint. Thus, we have considered 5-fold cross validation for our experimentation. We have reported the average of 5-fold cross-validation in our analysis.

5. Findings

This section reports the accuracy of our proposed classification method by using disease-specific word vectors. We have calculated the accuracy by comparing the predicted class (of the test data) with the predefined class of the annotated test data. In this context of multi-label classification, our accuracy reports the percentage of labels predicted for the test sample which exactly matches the corresponding set of annotated labels. To compare the accuracy of our domain-specific *word vectors* in identifying relevant tweets, we have considered publicly available pre-trained *word2vec* (contrived from Google news) and three generic pre-trained *Glove models* (prepared by Stanford NLP Group) as follows:

1 word2vec (on the basis of Google news.300)
2 Glove.840B300d (i.e., 840B tokens, 2.2 M vocab, cased, 300d)
3 Glove.6B50d (i.e., 6B tokens, 400 K vocab, uncased, 50d)
4 Glove.Twitter.27B50d (i.e., 2B tweets, 27B tokens, 1.2 M vocab, uncased, 50d)

Tables 4 and 5 report the classification accuracies of Ebola and Zika related tweets, respectively, of our word vectors (by using tweet corpora and scholarly abstracts), and compares the same with the accuracy of other pre-trained word vectors and their TF-IDF variants. As mentioned, our reported results are on the basis of extremely randomized tree-based ensemble method for supervised classification (Pedregosa et al., 2011). For the TF-IDF variants, we use the counts as the term-frequency. We observe that our word vectors (constructed from relatively small domain-specific Twitter and PubMed corpora) can match, and in some cases outperform, other pre-trained word vectors (constructed from voluminous generic input corpora like Google News or Wikipedia articles).

Our contribution to extant literature is on a number of fronts. We note that disease-specific *Word2Vec* has outperformed pre-trained *GloVe* (but not pre-trained *Word2Vec*) for Ebola outbreak. However, for the Zika outbreak our disease-specific *Word2Vec* has outperformed both pre-trained *GloVe* and pre-trained *Word2Vec*. On theoretical fronts, it suggests that a domain-specific word vector contrived from a mere 0.1 million tweets and can match the performance of word vectors contrived from billions and billions of tweets. Thus, for efficient text classification, quality and relevance of input corpora are more important than volume of the input corpora. This finding can have practical relevance for controlling outbreaks where getting voluminous quality input corpora is a constraint.

[1] *Word2Vec* implementation: https://radimrehurek.com/gensim/models/word2vec.html.

Table 4
Classification accuracies of Ebola-related tweets.

Word2Vec (GoogleNews.300)	0.78		
Word2Vec (GoogleNews.300)_TFIDF	0.72		
Glove.840B300d	0.75		
Glove.840B300d_TFIDF	0.72		
Glove.6B50d	0.70		
Glove.6B50d_TFIDF	0.64		
Glove.Twitter.27B50d	0.61		
Glove.Twitter.27B50d_TFIDF	0.58		
CBOW (D = 50)	*W = 3*	*W = 5*	*W = 8*
Word2Vec (tweets)	0.58	0.62	0.63
Word2Vec_TFIDF (tweets)	0.60	0.64	0.62
Word2Vec (PubMed)	0.46	0.48	0.49
Word2Vec_TFIDF (PubMed)	0.47	0.47	0.46
Skip-gram (D = 50)	*W = 5*	*W = 10*	*W = 15*
Word2Vec (tweets)	0.73	0.76	0.75
Word2Vec_TFIDF (tweets)	0.69	0.72	0.72
Word2Vec (PubMed)	0.64	0.69	0.68
Word2Vec_TFIDF (PubMed)	0.62	0.62	0.62
CBOW (D = 100)	*W = 3*	*W = 5*	*W = 8*
Word2Vec (tweets)	0.61	0.64	0.63
Word2Vec_TFIDF (tweets)	0.60	0.64	0.62
Word2Vec (PubMed)	0.47	0.47	0.50
Word2Vec_TFIDF (PubMed)	0.48	0.47	0.47
Skip-gram (D = 100)	*W = 5*	*W = 10*	*W = 15*
Word2Vec (tweets)	0.75	0.77	0.77
Word2Vec_TFIDF (tweets)	0.71	0.72	0.75
Word2Vec (PubMed)	0.69	0.71	0.71
Word2Vec_TFIDF (PubMed)	0.62	0.62	0.62
CBOW (D = 200)	*W = 3*	*W = 5*	*W = 8*
Word2Vec (tweets)	0.62	0.64	0.65
Word2Vec_TFIDF (tweets)	0.63	0.61	0.64
Word2Vec (PubMed)	0.47	0.47	0.49
Word2Vec_TFIDF (PubMed)	0.48	0.47	0.47
Skip-gram (D = 200)	*W = 5*	*W = 10*	*W = 15*
Word2Vec (tweets)	0.77	0.77	0.76
Word2Vec_TFIDF (tweets)	0.73	0.73	0.73
Word2Vec (PubMed)	0.71	0.73	0.73
Word2Vec_TFIDF (PubMed)	0.68	0.66	0.67

Figs. 2 and 3 report the comparison of various word vectors with changing labeled training samples. These figures indicate that our context-specific *Word2Vec* (contrived from Twitter corpus) are mostly outperforming other word vectors with better training. Interestingly, we have achieved the best accuracy for W = 5, D = 200, and Skip-gram model for both the outbreaks. Overall the accuracies of our domain-specific *Word2Vec* are at par with other pre-trained word vectors for Ebola and better for Zika outbreak. One possible reason might be the low volume of training and testing dataset for Ebola in comparison to Zika. Future studies need to investigate it further to make any conclusive comment. Intuitively word vectors contrived from a larger corpus is expected to perform better, but this might not be the case always. For instance, word vector of Zika contrived from fewer than 2000 abstracts delivered better results than the word vector of Ebola contrived from more than 4000 abstracts.

Furthermore, we observe that the word vectors from tweet corpus perform slightly better than word vectors from PubMed-based scholarly abstracts. However, it is worth noting that the accuracy of PubMed-based word vectors is also very high, and mostly at par, with other word vectors. More importantly, this study is an ex-post analysis where we have sufficient amount of Twitter data as input corpora. Voluminous Twitter data, even if we ignore the quality of the data, would be unavailable during the initial stages of the crisis. However, context specific biomedical literature is easily and always accessible as input corpora. Creating word vectors from the biomedical literature will be easy and time-effective. Thus, intuitively it can be argued that PubMed-based word vectors can outperform Twitter-based word vectors during the initial stages of an outbreak. The computational resource required for PubMed-based word vectors will be significantly lower in comparison to other pre-trained generic word vectors which considered millions of Wikipedia articles or billions of Google news. Thus, from the implementation perspective, PubMed based word vectors are viable cost-effective (in comparison to the computational resource required for pre-trained generic word vectors) and time-independent (in comparison to Twitter-based word vector) alternative during the outbreak.

On the contrary to short tweet feeds, biomedical literature comprises of long text. Long texts are better suited for word sense identification. Thus, in comparison to tweet feeds, extraction of semantic meaning and relation from the long biomedical text would be more meaningful. In a way, our study also demonstrates the possibility of knowledge transfer by using long text-based word vector for classifying short text. On practical fronts, this approach can be extrapolated to other outbreaks where a plethora of biomedical literature is easily available such as influenza-like illness which is a common and recurring health concern across the world. Twitter corpus-based word embedding is not the only option for classifying relevant and actionable tweets during the crisis. Scholarly texts

Table 5
Classification accuracies of Zika-related tweets.

Word2Vec (GoogleNews.300)	0.88		
Word2Vec (GoogleNews.300)_TFIDF	0.85		
Glove.840B300d	0.89		
Glove.840B300d_TFIDF	0.86		
Glove.6B50d	0.83		
Glove.6B50d_TFIDF	0.79		
Glove.Twitter.27B50d	0.77		
Glove.Twitter.27B50d_TFIDF	0.74		
CBOW (D = 50)	*W = 3*	*Win = 5*	*Win = 8*
Word2Vec (tweets)	0.89	0.89	0.90
Word2Vec_TFIDF (tweets)	0.87	0.87	0.87
Word2Vec (PubMed)	0.76	0.78	0.79
Word2Vec_TFIDF (PubMed)	0.75	0.77	0.77
Skip-gram (D = 50)	*W = 5*	*W = 10*	*W = 15*
Word2Vec (tweets)	0.91	0.91	0.91
Word2Vec_TFIDF (tweets)	0.87	0.88	0.89
Word2Vec (PubMed)	0.90	0.89	0.90
Word2Vec_TFIDF (PubMed)	0.87	0.85	0.85
CBOW (D = 100)	*W = 3*	*W = 5*	*W = 8*
Word2Vec (tweets)	0.89	0.90	0.90
Word2Vec_TFIDF (tweets)	0.87	0.88	0.87
Word2Vec (PubMed)	0.77	0.78	0.79
Word2Vec_TFIDF (PubMed)	0.77	0.76	0.78
Skip-gram (D = 100)	*W = 5*	*W = 10*	*W = 15*
Word2Vec (tweets)	0.92	0.90	0.91
Word2Vec_TFIDF (tweets)	0.87	0.87	0.88
Word2Vec (PubMed)	0.91	0.90	0.89
Word2Vec_TFIDF (PubMed)	0.88	0.87	0.87
CBOW (D = 200)	*W = 3*	*W = 5*	*W = 8*
Word2Vec (Tweets)	0.89	0.90	0.90
Word2Vec_TFIDF (tweets)	0.88	0.87	0.88
Word2Vec (PubMed)	0.76	0.77	0.77
Word2Vec_TFIDF (PubMed)	0.75	0.75	0.76
Skip-gram (D = 200)	*W = 5*	*W = 10*	*W = 15*
Word2Vec (tweets)	0.92	0.91	0.91
Word2Vec_TFIDF (tweets)	0.86	0.88	0.88
Word2Vec (PubMed)	0.91	0.90	0.90
Word2Vec_TFIDF (PubMed)	0.88	0.87	0.87

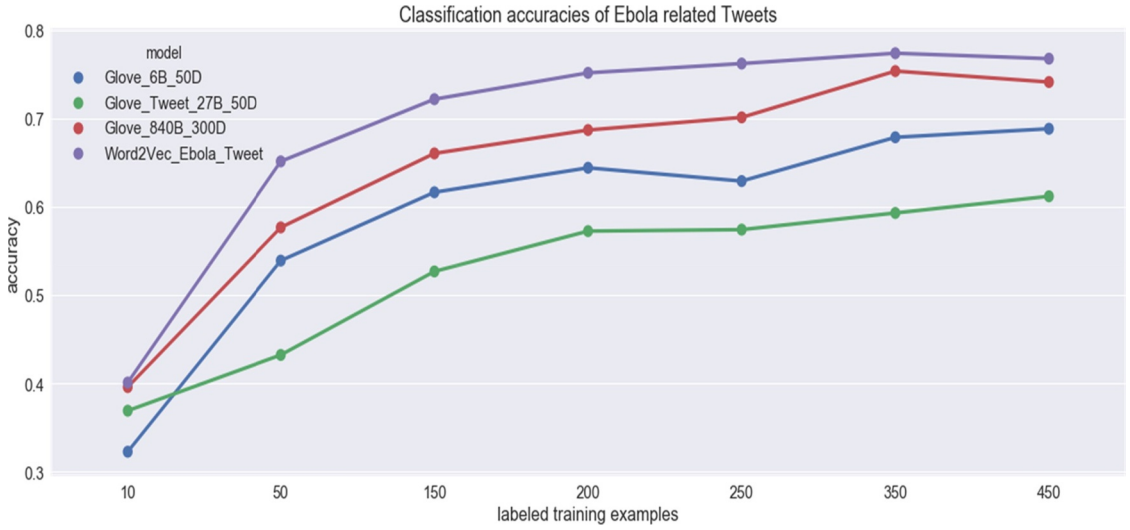

Fig. 2. Classification accuracies of Ebola-related tweets.
(Input corpus = tweets; W = 5; D = 200; Skip-gram).

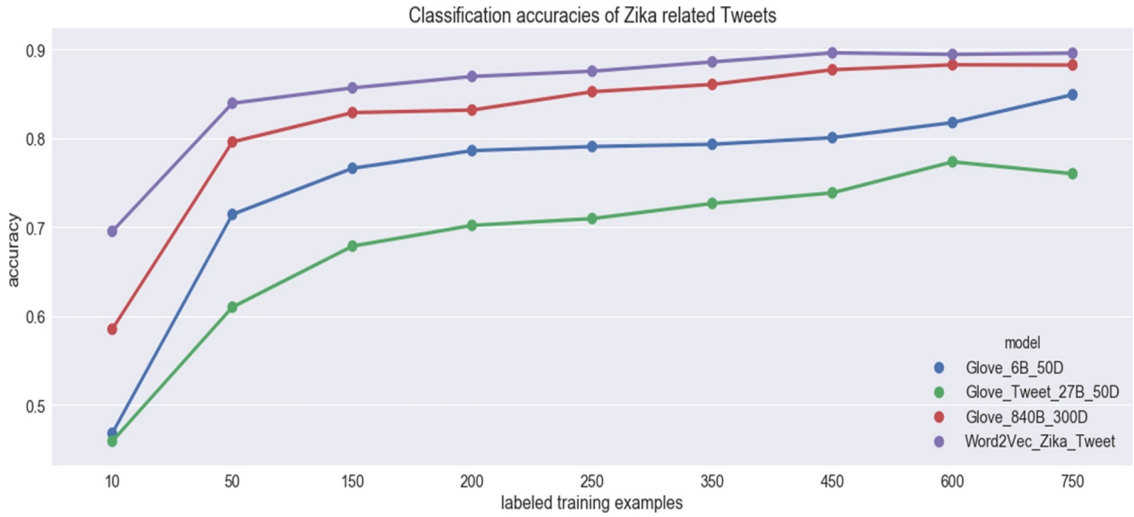

Fig. 3. Classification accuracies of Zika-related tweets.
(Input corpus = tweets; W = 5; D = 200; Skip-gram).

address the data scarcity concerns in the medical domain. Robustness of our results, across various hyper-parameters setting, re-confirms the validity of our approach.

On theoretical fronts, we observe that the Skip-gram models show better results than CBOW in our analysis. This pattern remains consistent for both the epidemics. Our findings are consistent with prior studies in biomedical contexts such as Chiu et al. (2016) and TH et al. (2015). Chiu et al. (2016) argue that in the CBOW 'the prediction back-propagated averaged over all context words' whereas in the Skip-gram 'the prediction error of each context word back-propagated to the target word,' and this is a possible explanation of this pattern.

6. Conclusion

Nowadays researchers are employing word vectors for 'a variety of linguistic tasks with considerable accuracy' but this approach might 'fail to generate satisfactory embeddings for highly specific domains such as healthcare where uncovering the relationships with respect to domain-specific words is of greater importance than the non-domain ones' (Ghosh et al., 2016). Moreover, Yang et al. (2016) noted that 'the effect of the configuration used to train and generate the word embedding on the classification performance has not been studied in the existing literature.' Our paper has attempted to address these research gaps. We have investigated how does the accuracy of word vectors change in response to input corpora (Twitter-based corpus vis-à-vis PubMed abstracts), model architectures (CBOW vis-à-vis Skip-gram) and hyper-parameter settings (dimension and context window size) in the context of infectious disease outbreaks. Overall our study suggests that domain-specific input corpora can mostly outperform generic pre-trained word vectors, Skip-gram accuracies are better than CBOW, and higher dimensions yield better accuracy. As we mentioned, the accuracy of the Twitter input corpus is slightly better than PubMed abstracts as input corpus. For instance, the accuracy (for classification of Zika related tweets) of the Twitter corpus is 0.92 whereas the accuracy of PubMed corpus is 0.91 (W = 5, D = 200) (see Table 5). This finding is not counterintuitive because Yang et al. (2016) noted that 'the background data type should align with the Twitter classification data set to achieve a better performance.' However, the difference between 0.91 and 0.92 is extremely narrow. This paper employed extra tree classifier in the last level of the model. However, future studies can consider advanced natural language processing models such as recurrent neural networks (RNN) or long short-term memory (LSTM) based models for classification to do a comparative evaluation between Twitter corpus and PubMed corpus.

This study portrays that during the initial stages of the epidemic, when collecting and aggregating Twitter corpus would be challenging, PubMed can be a reliable alternate option for immediate relief operations. It is quite possible that a combined corpus might be more accurate than our approaches. Future studies need to probe this further. We also observe that the accuracies of Zika vector models are consistently better than Ebola vector models. Probably one reason might be the size of our gold standards. However, it is worth remembering that 'there is not a single machine learning model that can work for all kinds of datasets (Zhang, Zhao, & LeCun, 2015). To sum up, our study demonstrates the strength of cost-effective domain-specific word vector models by comparing it against pre-trained generic word vectors (contrived from voluminous input corpora) in the context of two infectious disease outbreaks. Exhaustive experimentation confirms the robustness of our findings.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ipm.2018.10.010.

References

Ashktorab, Z., Brown, C., Nandi, M., & Culotta, A. (2014). Tweedr: Mining twitter to inform disaster response. *Proceedings of ISCRAM* (pp. 354–358). .

Cambria, E., Poria, S., Gelbukh, A., & Thelwall, M. (2017). Sentiment analysis is a big suitcase. *IEEE Intelligent Systems, 32*(6), 74–80.

Cambria, E., Mazzocco, T., Hussain, A., & Eckl, C. (2011). Sentic Medoids: Organizing Affective Common Sense Knowledge in a Multi-Dimensional Vector Space. *LNCS, 6677*, 601–610.

Cambria, E., Hussain, A., Durrani, T., Havasi, C., Eckl, C., & Munro, J. (2010a). Sentic Computing for Patient Centered Applications. *Proceedings of ICSP* (pp. 1279–1282). .

Cambria, E., Hussain, A., Durrani, T., Havasi, C., & Eckl, C. (2010b). SenticSpace: Visualizing Opinions and Sentiments in a Multi-Dimensional Vector Space. *LNAI, 6279*, 385–393.

Chan, E. H., Sahai, V., Conrad, C., & Brownstein, J. S. (2011). Using web search query data to monitor dengue epidemics: A new model for neglected tropical disease surveillance. *PLOS Neglected Tropical Diseases, 5*(5), e1206. https://doi.org/10.1371/journal.pntd.0001206.

Chew, C., & Eysenbach, G. (2010). Pandemics in the age of Twitter: Content analysis of Tweets during the 2009 H1N1 outbreak. *PloS One, 5*(11), e14118. https://doi.org/10.1371/journal.pone.0014118.

Chiu, B., Crichton, G., Korhonen, A., & Pyysalo, S. (2016). How to train good word embeddings for biomedical NLP. *Proceedings of BioNLP* (pp. 166). .

Choi, E., Bahadori, M. T., Searles, E., Coffey, C., Thompson, M., Bost, J., et al. (2016). Multi-layer representation learning for medical concepts. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1495–1504). ACM.

Culotta, A. (2010). Towards detecting influenza epidemics by analyzing Twitter messages. *Proceedings of the first workshop on Social Media Analytics* (pp. 115–122). ACM.

Dietterich, T. G. (2000). Ensemble methods in machine learning. *International workshop on multiple classifier systems* (pp. 1–15). Berlin, Heidelberg: Springer.

Fernández-Luque, L., & Bau, T. (2015). Health and social media perfect storm of information. *Healthcare Informatics Research, 21*(2), 67–73.

Fernández-Reyes, F. C., Hermosillo-Valadez, J., & Montes-y-Gómez, M. (2018). A Prospect-Guided global query expansion strategy using word embeddings. *Information Processing & Management, 54*(1), 1–13.

Gao, H., Barbier, G., & Goolsby, R. (2011). Harnessing the crowdsourcing power of social media for disaster relief. *IEEE Intelligent Systems, 26*(3), 10–14.

Ghosh, S., Chakraborty, P., Cohn, E., Brownstein, J. S., & Ramakrishnan, N. (2016). Characterizing diseases from unstructured text: A vocabulary driven word2vec approach. *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management (CIKM)* (pp. 1129–1138). .

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature, 457*(7232), 1012–1014.

Glorot, X., Bordes, A., & Bengio, Y. (2011). Domain adaptation for large-scale sentiment classification: A deep learning approach. *Proceedings of the 28th international conference on machine learning (ICML)* (pp. 513–520). .

Goldberg, Y., & Levy, O. (2014). word2vec Explained: Deriving Mikolov et al.'s negative-sampling word-embedding method. arXiv preprint arXiv:1402.3722.

Hickmann, K. S., Fairchild, G., Priedhorsky, R., Generous, N., Hyman, J. M., Deshpande, A., et al. (2015). Forecasting the 2013–2014 influenza season using Wikipedia. *PLoS Computational Biology, 11*(5), e1004239. https://doi.org/10.1371/journal.pcbi.1004239.

Imran, M., Elbassuoni, S., Castillo, C., Diaz, F., & Meier, P. (2013). Practical extraction of disaster-relevant information from social media. *Proceedings of the 22nd International Conference on World Wide Web (WWW)* (pp. 1021–1024). ACM.

Jin, F., Wang, W., Zhao, L., Dougherty, E., Cao, Y., Lu, C. T., et al. (2014). Misinformation propagation in the age of twitter. *Computer, 47*(12), 90–94.

Kalyanam, J., Velupillai, S., Doan, S., Conway, M., & Lanckriet, G. (2015). Facts and Fabrications about Ebola: A Twitter Based Study. *Proceedings of the ACM KDD Workshop on Connected Health in Big Data Era*.

Khatua, A., & Khatua, A. (2016). Immediate and long-term effects of 2016 Zika Outbreak: A Twitter-based study. *Proceedings of 18th International Conference on e-Health Networking, Applications and Services (Healthcom)* (pp. 1–6). IEEE.

Lampos, V., Zou, B., & Cox, I. J. (2017). Enhancing feature selection using word embeddings: The case of flu surveillance. *Proceedings of the 26th International Conference on World Wide Web (WWW)* (pp. 695–704). .

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google Flu: Traps in big data analysis. *Science, 343*(6176), 1203–1205.

Lilleberg, J., Zhu, Y., & Zhang, Y. (2015). Support vector machines and word2vec for text classification with semantic features. *Cognitive Informatics & Cognitive Computing (ICCI* CC), 2015 IEEE 14th International Conference on* (pp. 136–140). IEEE.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013a). Efficient estimation of word representations in vector space. Available at arXiv preprint arXiv:1301.3781.

Mikolov, T., Yih, W. T., & Zweig, G. (2013b). Linguistic regularities in continuous space word representations. *Proceedings of HLT-NAACl* (pp. 746–751). .

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013c). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems* (pp. 3111–3119). .

Miller, M., Banerjee, D., Muppalla, R., Romine, D., & Sheth, A. (2017). What Are People Tweeting About Zika? An Exploratory Study Concerning Its Symptoms, Treatment, Transmission, and Prevention. *JMIR Public Health Surveillance, 3*(2), e38. https://doi.org/10.2196/publichealth.7157.

Minarro-Giménez, J. A., Marin-Alonso, O., & Samwald, M. (2014). Exploring the application of deep learning techniques on medical text corpora. *Studies in Health Technology and Informatics, 205*, 584–588.

Oyeyemi, S. O., Gabarron, E., & Wynn, R. (2014). Ebola, Twitter, and misinformation: A dangerous combination? *BMJ, 349*, g6178. https://doi.org/10.1136/bmj.g6178.

Pakhomov, S., McInnes, B., Adam, T., Liu, Y., Pedersen, T., & Melton, G. B. (2010). Semantic similarity and relatedness between clinical terms: An experimental study. *AMIA Annual Symposium Proceedings, 2010*, 572–576.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of EMNLP* (pp. 532–1543). .

Poria, S., Gelbukh, A., Cambria, E., Yang, P., Hussain, A., & Durrani, T. (2012). Merging SenticNet and WordNet-Affect emotion lists for sentiment analysis. *Proceedings of ICSP* (pp. 1251–1255). .

Rong, X. (2014). Word2Vec parameter learning explained. Available at arXiv preprint arXiv:1411.2738.

Rudra, K., Sharma, A., Ganguly, N., & Imran, M. (2018). Classifying and Summarizing Information from Microblogs During Epidemics. *Information Systems Frontiers, 20*(5), 933–948.

TH, M., Sahu, S., & Anand, A. (2015). Evaluating distributed word representations for capturing semantics of biomedical concepts. *Proceedings of BioNLP. 15. Proceedings of BioNLP* (pp. 158–163).

Turner, C. A., Jacobs, A. D., Marques, C. K., Oates, J. C., Kamen, D. L., Anderson, P. E., et al. (2017). Word2Vec inversion and traditional text classifiers for phenotyping lupus. *BMC Medical Informatics and Decision Making, 17*(1), 126. https://doi.org/10.1186/s12911-017-0518-1.

Vieweg, S., Hughes, A. L., Starbird, K., & Palen, L. (2010). Microblogging during two natural hazards events: what Twitter may contribute to situational awareness. In *Proceedings of the SIGCHI conference on human factors in computing systems*, ACM, 1079-1088.

Wang, X., Kumar, D., Thome, N., Cord, M., & Precioso, F. (2015, June). Recipe recognition with large multimodal food dataset. *Multimedia & Expo Workshops (ICMEW), 2015 IEEE International Conference on* (pp. 1–6). IEEE.

Yang, X., Macdonald, C., & Ounis, I. (2016). Using word embeddings in twitter election classification. Available at arXiv preprint arXiv:1606.07006.

Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. *Advances in Neural Information Processing Systems* (pp. 649–657). .

Zhu, Y., Yan, E., & Wang, F. (2017). Semantic relatedness and similarity of biomedical terms: examining the effects of recency, size, and section of biomedical publications on the performance of word2vec. *BMC Medical Informatics and Decision Making, 17*(1), 95. https://doi.org/10.1186/s12911-017-0498-1.