International Journal of Data Science and Analytics
https://doi.org/10.1007/s41060-022-00344-x

REGULAR PAPER

Domain-specific text dictionaries for text analytics

Andrea Villanes[1] · Christopher G. Healey[1]

Received: 16 December 2021 / Accepted: 24 June 2022

Abstract
We investigate the use of sentiment dictionaries to estimate sentiment for large document collections. Our goal in this paper is a semiautomatic method for extending a general sentiment dictionary for a specific target domain in a way that minimizes manual effort. General sentiment dictionaries may not contain terms important to the target domain or may score terms in ways that are inappropriate for the target domain. We combine statistical term identification and term evaluation using Amazon Mechanical Turk to extend the EmoLex sentiment dictionary to a domain-specific study of dengue fever. The same approach can be applied to any term-based sentiment dictionary or target domain. We explain how terms are identified for inclusion or re-evaluation and how Mechanical Turk generates scores for the identified terms. Examples are provided that compare EmoLex sentiment estimates before and after it is extended. We conclude by describing how our sentiment estimates can be integrated into an epidemiology surveillance system that includes sentiment visualization and discussing the strengths and limitations of our work.

Keywords Dengue fever · Text analytics · Sentiment

1 Introduction

This paper investigates the use of sentiment dictionaries for estimating sentiment in text. Together with machine learning approaches, sentiment dictionaries are a common method for assigning sentiment to text. The simplest approaches use polarity to classify text as positive–negative or positive–neutral–negative. More sophisticated methods use emotional dimensions from psychology to characterize the sentiment implied by a text block [1–3].

One important advantage of sentiment dictionaries is they are unsupervised: No training set or labeled examples are required to use them. General sentiment dictionaries also have a number of limitations, however.

1. *Domain context* We cannot specialize a term's emotional scores based on a target domain.
2. *Missing terms* We cannot score terms a dictionary does not contain.
3. *Term independence* We cannot derive context from neighboring terms, e.g., "I am **happy**" versus "I am **not happy**."
4. *Term ambiguity* We cannot differentiate between homonyms, e.g., "I lie down" versus "I lie often."

This paper focuses on the generality of a dictionary's entries, which address the first two limitations listed above. By design, sentiment dictionaries are built to function over a wide range of text domains. This approach maximizes their relevance, but it also means that domain-specific sentiment is not available. The lack of availability can lead to inaccurate estimates when terms with a unique emotional affect for a given domain are scored.

We propose a semiautomatic method to modify and extend a sentiment dictionary for a user-chosen domain. First, two types of terms are identified: *unique terms* that are important to the domain but not present in a dictionary and *common terms* that exist in the dictionary but possibly with incorrect sentiment scores. Statistical analysis is used to select the unique and common terms to evaluate, which is typically significantly smaller than the overall size of the original dictionary. Amazon Mechanical Turk (MTurk) is used to obtain new scores that are integrated back into the original dictio-

Andrea Villanes and Christopher G. Healey have contributed equally to this work

✉ Christopher G. Healey
healey@ncsu.edu

[1] Department of Computer Science and Institute for Advanced Analytics, North Carolina State University, 890 Oval Drive, Raleigh, NC 27695-8206, USA

Published online: 11 July 2022

nary. The result is a dictionary that more accurately estimates sentiment for terms important to the domain under evaluation.

To demonstrate our approach, we use dengue fever as our example domain [4], Plutchik's sentiment model for emotional dimensions [5], and Mohammad's EmoLex dictionary for sentiment estimation [6]. The same approach can be used for any domain, sentiment model, and dictionary, as long as sufficient text documents from the domain are available. We conclude by describing how our sentiment estimates can be integrated into an epidemiology surveillance system, a critical tool for tracking disease onset and progression in regions where up-to-date information is unavailable. Our use case generates modified Rose charts to visualize text sentiment over positive and negative valence, and Plutchik's eight emotional dimensions. Finally, we enumerate limitations of our system for future work.

2 Background

Our goals are to identify terms in a general sentiment dictionary that require: (1) addition to the dictionary or (2) re-evaluation of their sentiment in the context of the target domain. We begin by describing the general area of sentiment analysis, with a focus on past and current natural language processing (NLP) and lexicon-based approaches. We identify potential limitations in each area based on our goals and explain our choice of lexicon-based sentiment dictionaries due to their unsupervised nature, which frees us from finding or building a pre-labeled sentiment dataset for training.

2.1 Sentiment analysis

Sentiment analysis is an active research area in natural language processing (NLP), information retrieval (IR), and machine learning (ML). Two common analysis methods are: (1) supervised, using a training set to build emotion estimation models, and (2) unsupervised, where raw text is converted directly into scores along emotional dimensions [7–10].

Analysis is often built on psychological models of emotion that use orthogonal dimensions to describe emotional affect. For example, Russell defined three dimensions pleasure (or valence), arousal, and dominance—the PAD model—to represent emotion [11,12] (Fig. 1a). Plutchik's four-dimensional model of joy–sadness, anger–fear, trust–disgust, and anticipation–surprise uses a color wheel to represent basic emotions: hue for dimension endpoints (eight hues) and saturation for emotional intensity (weak saturation for low intensity to strong for high, Fig. 1b) [5].

2.2 Sentiment estimation

In the area of supervised NLP approaches, preprocessing has been applied prior to sentiment analysis. Pang and Lee calculated subjectivity weights for sentences using ML, producing a graph of sentence nodes and subjectivity-weighted edges [13]. A minimum graph cut is used to separate objective and subjective sentences. Pang et al. also compared Naïve Bayes, maximum entropy, and support vector machines (SVMs) for classifying movie reviews as positive or negative [14]. Unigrams performed best using SVM. Augmenting the training set with intuitive extensions like bigrams, term frequencies, part of speech tagging, and document position information did not improve performance. Turney rated online reviews as positive or negative using pointwise mutual information to generate statistical dependence between review phrases and the anchor words "excellent" and "poor" [15].

Several pre-built sentiment analysis libraries are available [16,17]. For example, in Python, the Natural Language Toolkit's Valence Aware Dictionary and Sentiment Reasoner (NLTK VADER) scores text blocks both for polarity (negative, neutral, and positive) and overall sentiment (compound). Textblob includes a sentiment analysis engine among other common NLP algorithms (part-of-speech tagging, noun phrase extraction, and translation). Textblob returns a sentiment polarity score on the range $[-1, \ldots, 1]$ and a subjectivity score on the range $[0, 1]$. Finally, Flair uses a pre-trained word embedding model to perform sentiment analysis. Although slower than VADER or Textblob, tests suggest that Flair produces more accurate sentiment scores when compared to star ratings for product reviews. Other comparisons of VADER versus Textblob versus Flair exist, for example, on the CIA World Factbook [18].

These methods are simple to use and perform reasonably in a generalized environment. Two issues for our work are how to optimize the libraries for a target domain and how to redefine their polarity output for a more sophisticated emotional model. Pre- or post-processing may be able to handle the first issue, although identifying the terms to target would still need a method like the one proposed in our paper. Extension to different emotional models is more challenging and could require changes to the libraries themselves to complete.

Previous work has addressed the issue of domain-specific sentiment. Li et al. proposed the Hierarchical Attention Transfer Network (HATN) [19]. HATN takes sentiment-labeled examples from a source domain, then builds a model based on *pivots* (domain-shared sentiment terms) and *non-pivots* (domain-specific sentiment terms) to estimate sentiment for a target domain without an available training dataset. The intuition is that non-pivot words mirror the sentiment of neighboring pivot terms. Our approach is similar; however, we use domain-specific and general text to identify non-pivot terms, then employ Amazon Mechanical

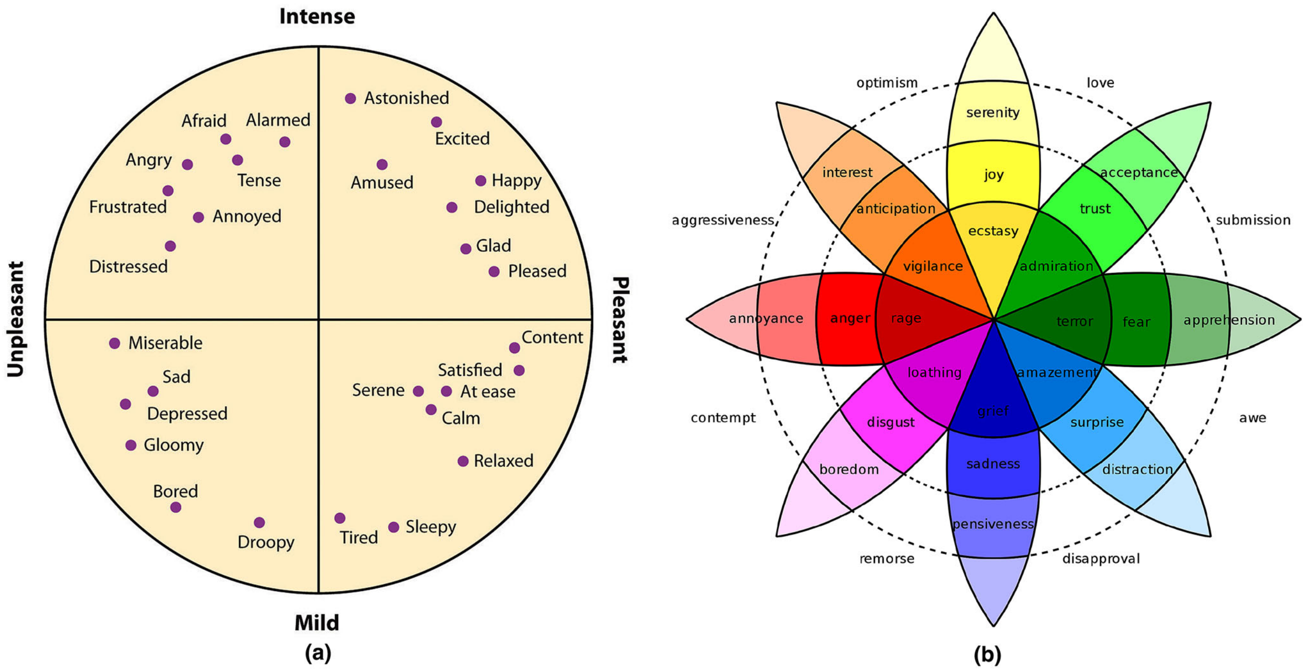

Fig. 1 Emotional models: **a** Russell's emotional circumplex, pleasure (valence) on the horizontal axis, arousal on the vertical axis; **b** Plutchik's emotional model, anger–fear on the horizontal axis, joy–sadness on the vertical axis, trust–disgust on the right-diagonal axis, and anticipation–surprise on the left-diagonal axis

Turk to assign sentiment to these terms. Zhang et al. use a dynamic re-weighting of the BERT language model to determine sentiment for the aspect of terms in a target sentence [20]. Aspect-modified terms are identified, then re-weighted using attention and multilayer deep learning. This is different from our approach, which does not extend an existing language model, but instead modifies a term-based sentiment dictionary to better represent sentiment for a target domain.

More recently, the areas of deep learning and deep neural networks (DL and DNNs) have been applied to NLP problems, including sentiment analysis, with great success. Initial work focused on recurrent neural networks, often augmented with long-short term memory (RNNs and LSTMs), since the sequence-based nature of RNNs seemed well suited to sequences of text tokens. Recently, RNN and LSTM models have been superseded by deep learning transformers.

A current and well-known DL approach is bidirectional encoder representations from transformers (BERT). BERT is a pre-trained language model developed by Google [21]. Built on the transformer-based deep learning model, BERT connects every output to every input, with weights dynamically calculated based on a model of *attention*, avoiding the RNN and LSTM issues of recognizing context over large text sequences. BERT examines the entire text sequence in both directions using bidirectionality. BERT was trained using two related tasks: (1) masked language models that hide a word and ask the model to predict the word based on its surrounding context, and (2) next sentence prediction where a probability for two sentences having a logical, sequential connection is calculated. BERT can be fine-tuned for different tasks, including sentiment analysis. A common approach is to use TensorFlow and a review database with star ratings like IMDB or Amazon to predict sentiment polarity. Once BERT is extended in this way, it can be applied to unlabeled text to estimate its sentiment.

Another well-known NLP system is Generative Pre-trained Transformer 3 (GPT-3), a commercial language model developed by OpenAI [22]. Similar to BERT, GPT-3 provides a pre-trained language model for addressing NLP tasks, including sentiment analysis. A key difference is that BERT uses bidirectional analysis, while GPT-3 uses autoregression. Another critical advantage of GPT-3 is its ability to use a few-shot learning process, allowing, in theory, the customization of sentiments for a target domain with a smaller training set. A potential drawback of GPT-3 is that it is not open source, although this can depend on the intended use case. Similar to BERT, GPT-3's pre-trained model can be extended to estimate sentiment, although this can often be done with fewer training samples.

If we view BERT as focusing on pre-training the encoding step in a deep learning transformer and GPT-3 as focusing on pre-training the decoding step, an obvious idea is to pre-train a complete encode–decode architecture. Masked sequence to sequence (MASS) and BART employ this approach, claiming to produce generalizations of BERT and GPT [23,24]. MASS masks out k consecutive tokens in the input sequence then attempts to predict those tokens in the output sequence. BART introduces noise into the input sequence (token masking, token deletion, token infilling, sentence shuffling, and document rotation) to generate "noisy" input for the encoder, then applies an autoregressive decoder to try to remove the noise and reconstruct the original input. Since MASS and BART and extensions of BERT and GPT-3, they can also be extended to estimate sentiment in similar ways.

Despite their enormous power, general purpose deep learning-based NLP models struggle with the lack of domain focus. For context, we fine-tuned BERT on IMDB reviews, then scored the six dengue sentences in upcoming Table 7. We used the HuggingFace library with parameters recommended by the original paper's authors: a batch size of 32, an Adam learning rate of 5×10^{-5}, three epochs, an ϵ of 1×10^{-8} to avoid division by zero, and a one-layer feed-forward classifier. BERT reported all six as positive, whereas our approach more accurately identifies a finer-grained range of positive and negative polarities. Training on a dengue-specific corpus would yield different results. This point highlights one of the potential drawbacks of BERT, however, since no pre-tagged sentiment training set exists for dengue text. Our approach to augment existing sentiment dictionaries through an explicit trade-off between domain relevance and user effort might be integrated into BERT-, BART- or GPT-based sentiment analysis as a post-processing step. The sentiment categories would need to be increased for a model like Plutchik's, although probability scores from a deep learning model for each category could act as proxies for the "amount" of a given sentiment type contained in the text.

2.3 Sentiment dictionaries

A common unsupervised approach employs sentiment dictionaries. Terms appear as keys, but each term is associated with one or more emotional dimension scores rather than definitions. POMS-ex (Profile of Mood States) is a 793-term dictionary designed to measure emotion on six dimensions: tension–anxiety, depression-dejection, anger–hostility, fatigue–inertia, vigor–activity, and confusion–bewilderment [25]. Affective Norms for English Words (ANEW) used the PAD model to score 1033 emotion-carrying terms along each dimension using a nine-point scale [26]. Mohammad and Turney created EmoLex from 14,182 nouns, verbs, adjectives, and adverbs using Plutchik's four emotional dimensions joy–sadness, anger–fear, trust–disgust, and anticipation–surprise [6]. Other dictionaries also exist: SentiStrength, built from MySpace comments [27]; Linguistic Inquiry and Word Count (LWIC), a dictionary that classifies terms as positive, negative, or neutral [28], and SentiWordNet, built from the well know WordNet synset dictionary [29]. More recently, researchers have applied Amazon Mechanical Turk to assign scores for emotional dimensions to large dictionaries. Warriner extended the original ANEW dictionary to approximately 13,000 terms [30] using MTurk to obtain PAD scores and compared results to the original ANEW scores for validation.

In recent years visualizing sentiment has received significant attention as part of the general text visualization area. Kucher et al. provide an overview of recent sentiment visualization techniques [3]. Cao et al. developed Whisper to monitor the spatiotemporal diffusion of social media information. Sentiment polarity was visualized using a sunflower metaphor to identify influencers and geolocated groups receiving and spreading information [31]. SocialHelix followed, visualizing and tracking social media topics as they form and their sentiment diverges using a DNA-like representation [32]. Wu et al. presented opinion propagation in Twitter using a combination of streamgraphs and Sankey graphs [33]. Liu et al. linked primary and secondary text using semantic lexical matching. The results are presented in a dashboard containing topic keywords, concept clusters, and a causality timeline [34]. El-Assadi et al. visualized multiparty conversation behavior at the topic level with ConToVi [35]. They also extracted conversation threads from large online conversation spaces using a combination of supervised and unsupervised machine learning algorithms [36]. Hoque and Carenini implemented ConVis and MultiConVis, an ML, NLP, and visual analytic system to explore blog conversations [37,38]. Mohammad et al. extracted stance and sentiment in tweets using a labeled database, with results visualized using treemaps, bar graphs, and heatmaps [39]. Kucher et al. identified stance and sentiment polarity in social media text, then used similarity over these properties to visualize analysis of collections of topic–data source streams [40]. Wei et al. proposed TIARA, a system to extract topics that are visualized in an annotated streamgraph [41]. Dörk et al. use a construct called Topic Streams, a streamgraph approach to monitoring topics in a large online conversation environment over time [42].

Despite this significant progress, numerous challenges in sentiment estimation continue to exist: more subtle text cues (e.g., sarcasm, irony, humor, or metaphors), a writer's emotion versus what they write (e.g., an author evoking a particular emotional affect), emotion toward different aspects of an entity, stance (i.e., the opinion on a topic), or cross-cultural and domain differences (e.g., "alcohol" can be evaluated differently depending on the underlying culture) [8,9,43].

In the end, we chose to modify an existing sentiment dictionary to customize it for a target domain. Our motivation for this approach is the advantage of avoiding the need for a pre-labeled training set. This technique does not preclude manual work, however, since existing terms may need to be evaluated to update their sentiment for the target domain, and terms important to the domain may need to be added to the dictionary.

2.4 Dengue fever dictionary

Dengue is a mosquito-borne viral disease transmitted to humans through infected Aedes mosquitoes, a tropical and subtropical species found throughout the world. Common symptoms of dengue include persistent vomiting, fluid accumulation, lethargy, rash, and pain. To date, there is no available vaccine for dengue [44]. Dengue spread rapidly during the twentieth century to infect more than 300 million people in 2010 [45]. One in three people live among mosquitoes that transmit the dengue virus, yet there remain major uncertainties over the burden of dengue [46–49]. New, improved methods for assessing this burden are in critical demand [50].

Communicable diseases remain among the leading mortality causes in many countries, particularly in Asia and Africa [51]. In 2010, of the 52.8 million deaths globally, 24.9% were due to communicable, maternal, neonatal, and nutritional causes. 76% of premature mortality in sub-Saharan Africa in 2010 was due to the same causes [51]. Combating communicable diseases depends on surveillance, preventive measures, outbreak investigation, and the establishment of control mechanisms [52]. Unfortunately, data from surveillance systems are often delayed, and reporting is inaccurate, making it difficult to use such data for the detection of outbreaks [53–57]. It is estimated that only 35% of communicable disease cases are reported to national health departments [58–60].

We integrated our sentiment analysis results into a surveillance system for dengue outbreaks in India. Although India is one of the few countries that publish government data on outbreak statistics, information is neither timely nor accurate. The most recent Indian government reports on dengue are for 2014. A recent study by Shepard et al. identified underreporting of cases by 282× for one district in India [4]. For example, a single hospital in Lucknow in the state of Uttar Pradesh reported 216 dengue cases [61], where the government reported only 200 cases for the entire state of Uttar Pradesh.

3 Methods

Here, we discuss how to identify terms to re-evaluate or add to a sentiment dictionary using a statistically driven approach built to minimize manual effort.

3.1 Domain-specific terms

Sentiment dictionaries like POMS-ex, ANEW, and EmoLex are meant to be applied in a general context. By design, they are not built to focus on any particular topic. This decision broadens their relevance, but it also leads to the possibility of missing or incorrectly scored terms when used in a specific research domain. Our interest is in a semiautomatic method to update and extend a sentiment dictionary for a user-chosen research area of interest.

Our algorithm uses relative term frequency to identify terms that are “important” to a target domain. Consider a document collection D from the target domain. Individual terms t_i are identified, and their per-document frequencies $n^D_{t_i,j} = |t_i| \in d_j$ are calculated. Next, we construct a general document collection G using documents that are not specific to the target domain, drawn uniformly from the document space. For example, for our dengue fever research, D represents English language newspaper articles about dengue fever, and G represents an equivalent number of English language newspaper articles that do not discuss dengue, drawn uniformly from the Brandwatch[1] newspaper database that forms our document space. In our dengue research $|D| = 981,743$ total terms and $|G| = 922,897$ total terms. As with the target articles, we enumerate individual terms and term frequencies $n^G_{t_i,j} = |t_i| \in g_j$. Given D and G, we identify terms that fall into two categories.

1. *Unique terms* $t_i \notin G$ whose total frequency $N^D_{t_i} = \sum_j n^D_{t_i,j}$ exceeds a threshold value. t_i represent high-frequency terms that do not exist in EmoLex, so emotional scores for these terms must be obtained.
2. *Common terms* $t_i \in D, G$ where total document frequency $N^D_{t_i}$ is statistically significantly larger than total frequency $N^G_{t_i} = \sum_j n^G_{t_i,j}$. t_i represent high-frequency terms that are assumed to be important to the target domain and therefore could have emotional scores that are different from the term’s general use.

3.2 Significant frequency difference

Unique terms t_i are automatically flagged for evaluation. Any unique t_i with $N^D_{t_i} > 20$ is included. This cutoff was chosen: (1) to select terms only when a sufficient number of occurrences in the domain corpus were found and (2) to mirror term evaluation frequencies in other sentiment dictionaries like ANEW and POMS-ex. The required value of $N^D_{t_i}$ can

[1] https://www.brandwatch.com, formerly Crimson Hexagon, a subscription service that provides “insights from 100 million sources and 1.4 trillion posts”

Table 1 Contingency table for term t_i: $N^D_{t_i}$, $N^G_{t_i}$ and $N^D_{\bar{t}_i}$, $N^G_{\bar{t}_i}$ represent the frequency of t_i and not t_i in D and G, respectively, $|D|$ and $|G|$ represent total terms in D and G, respectively

	$\|t_i\|$	$\|\bar{t}_i\|$
D	$N^D_{t_i}$	$N^D_{\bar{t}_i} = \|D\| - N^D_{t_i}$
G	$N^G_{t_i}$	$N^G_{\bar{t}_i} = \|G\| - N^G_{t_i}$
Odds	$O^D_{t_i} = {}^{N^D_{t_i}} / {}_{\|D\|}$	$O^D_{\bar{t}_i} = {}^{N^D_{\bar{t}_i}} / {}_{\|D\|}$
	$O^G_{t_i} = {}^{N^G_{t_i}} / {}_{\|G\|}$	$O^G_{\bar{t}_i} = {}^{N^G_{\bar{t}_i}} / {}_{\|G\|}$

Table 2 Contingency table for term t_i=mosquito

	|mosquito|	$\|\overline{\text{mosquito}}\|$	Total
Dengue	7647	974,096	981,743
General	12	922,885	922,897
Odds	0.007853	0.992147	1.0
	0.000013	0.999987	1.0

be varied to increase or decrease the number of unique terms selected for evaluation.

Common terms are only re-evaluated if the frequency N^D_i is significantly higher than N^G_i. We use Fisher's exact test to calculate the probability of the null hypothesis: the odds ratio[2] of $t_i \in D$ with respect to $t_i \in G$ is 1 [62,63].

Consider a 2 × 2 contingency table (Table 1). Assuming a multinomial distribution of terms $\pi(t_i, D)$ and $\pi(t_i, G)$ in D and G, respectively, the null hypothesis can be stated as $H_0 : \pi(t_i, D) = \pi(t_i, G)$. The odds $O^D_{t_i}$ and $O^G_{t_i}$ represent the likelihood of t_i occurring in either newspaper collection.

The odds *ratio* represents the relative strength of the relationship between t_i in D versus t_i in G.

$$\theta_{t_i} = \frac{O^D_{t_i}}{O^G_{t_i}} \tag{1}$$

θ_{t_i} measures the likelihood of finding t_i in D versus finding it in G. If the odds of t_i are the same for both D and G then $\theta_{t_i} = 1$, the null hypothesis. Given fixed row totals $|D|$ and $|G|$, knowing $N^D_{t_i}$ determines the values $|D| - N^D_{t_i}$, $|G| - N^G_{t_i}$, and $N^G_{t_i}$ for the other three cells in the contingency table. Assuming independent binomial sampling, fixed row totals allow us to estimate the conditional distribution for both $\pi(t_i, D)$ and $\pi(t_i, G)$. The test of independence in a 2 × 2 table is now equivalent to testing for $\theta_{t_i} = 1$.

In terms of Fisher's test, H_0 represent the frequency of $t_i \in D$ as statistically equivalent to $t_i \in G$. To calculate this probability, Fisher uses the following formula.

$$\begin{aligned} p(t_i) &= \frac{\dbinom{|D|}{N^D_{t_i}}\dbinom{|G|}{N^G_{t_i}}}{\dbinom{N}{N_{t_i}}} \\ &= \frac{|D|!\ |G|!\ N_{t_i}!\ N_{\bar{t}_i}!}{N^D_{t_i}!\ (|D| - N^D_{t_i})!\ N^G_{t_i}!\ (|G| - N^G_{t_i})!} \end{aligned} \tag{2}$$

where $N_{t_i} = N^D_{t_i} + N^G_{t_i}$ and $N_{\bar{t}_i} = N^D_{\bar{t}_i} + N^G_{\bar{t}_i}$. Since most of the factorials in Fisher's formula are large, we approximate them with Sterling's formula.

$$\log n! \approx n \log n - n + \frac{1}{2} \log n + \log \sqrt{2\pi} \tag{3}$$

Consider a practical example for t_i=mosquito. Based on our document collection, the contingency table for mosquito is shown in Table 2. Equation 1 is used to compute $\theta_{\text{mosquito}} = {}^{0.0078503} / {}_{0.000013} = 603.87$, that is, t_i=mosquito is approximately 604 times more likely to occur in dengue newspaper articles versus general newspaper articles. Using Eqns 2 and 3 produce $p(\text{mosquito}) < 0.00001$. Not surprisingly, the term mosquito is significantly more likely to occur in dengue articles versus general articles.

Given the ability to compute p for each term in our common term set, we next defined a set of rules to determine when a common term required re-evaluation (Table 3). Terms with $p(t_i) < 0.00001$ and $\theta_{t_i} \geq 5$ were flagged for re-evaluation. These thresholds represent terms with a significantly higher odds ratio in dengue newspaper articles versus general newspaper articles. A total of 602 terms were identified for re-evaluation using these rules.

3.3 Term evaluation

A total of 850 terms (248 unique, 602 common) were evaluated for our dengue domain. Evaluation was performed with Amazon Mechanical Turk, the same crowd-sourcing platform used to construct EmoLex. MTurk acts as an online marketplace for experiment design and execution using Amazon's vast collection of *Turkers* or *workers*. Each worker completes Human Intelligence Tasks (HITs) that normally represent one trial in an experiment. Workers are paid based on the number of HITs they successfully complete.

We quickly realized that in order to obtain high-quality results, we needed to recruit MTurk "master Workers" [64]. Amazon defines a Master worker as one who "...consistently demonstrates a high degree of success in performing a wide range of HITs across a large number of Requesters." Master workers are more expensive and more difficult to recruit since they often "test" an experiment to ensure they will be paid properly before they fully commit to its HITs. The advantage

[2] The odds a particular outcome occurs given a particular exposure, versus the odds of the outcome absent the exposure.

Table 3 Rules for common term re-evaluation

Type	p	Odds Ratio	n	Re-evaluate
Not Different	$p \geq 0.01$	$0.11 < \theta_{t_i} < 9.4$	10, 022	No
Possibly Different	$0.00001 \leq p < 0.01$	$0.05 \leq \theta_{t_i} < 19.74$	2368	No
Different	$p < 0.00001$	$0 < \theta_{t_i} < 5$	2248	No
Different	$p < 0.00001$	$\theta_{t_i} \geq 5$	602	Yes

is that master workers produce much better results and will often complete many more HITs than a regular worker.

Our experiment task asks workers to identify the emotions associated with target terms in the context of dengue fever. We present twelve multiple-choice questions for each term: two filter questions (Table 4) and ten experiment questions. The first filter question presents the target term and four different words. The worker is asked which word is closest in meaning (synonym) to the target term. This testing ensures native and fluent English. The second filter question presents a short paragraph defining dengue fever then asks a multiple-choice question about dengue. This question ensures an understanding of the context for the experiment questions. Workers must answer both filter questions correctly to continue.

The remaining ten questions correspond to positive and negative valence and to Plutchik's eight emotions. Workers are asked to rate the term as having a Strong, Moderate, Weak, or No correspondence to the following ten properties.

1. In terms of dengue fever, how *positive* is the term?
2. How *negative* is the term?
3. How much is the term associated with *joy*?
4. How much is the term associated with *sadness*?
5. How much is the term associated with *anger*?
6. How much is the term associated with *fear*?
7. How much is the term associated with *trust*?
8. How much is the term associated with *disgust*?
9. How much is the term associated with *surprise*?
10. How much is the term associated with *anticipation*?

Answers to these questions provided an overall valence score and scores for each of Plutchik's eight emotional dimensions. The two filter questions plus the ten term questions formed one MTurk HIT.

4 Results

Questions in each HIT we presented were validated. We collected five independent HITs for each of the 850 evaluation terms, producing a total of 4250 HITs completed by 141 workers. Any incorrect filter question or unanswered question removed a HIT from the HIT set. Fifty-seven HITs failed filter question one, 27 HITs failed filter question two, and 40 term questions were unanswered. After removal, we were left with 4137 HITs representing 97.3% of the initial HIT set.

Next, HIT outliers were removed. Five independent workers evaluated each HIT's two valence and eight emotion questions for a total of 50 responses, 5 per emotion. We flagged responses outside the standard 1.5 IQR (interquartile range) as outliers. An individual worker's HIT can have at most ten outliers. We removed 138 HITs with four or more outliers, retaining 3999 HITs (94.1%). We define this HIT collection as our master HIT set. Within this set, 128 workers completed the 3999 HITs in a median and mean time of 40 and 112 s per HIT, respectively. A minimum of three HITs for every term was present in the master set, with a median and mean of 5 and 4.7 HITs, respectively. On average, a worker completed 31 HITs.

4.1 Analysis

Table 5 lists, for each of Plutchik's eight emotions and the two valence terms *Positive* and *Negative*, the percentage of annotations in the master set associated with the four intensity levels Strong, Moderate, Weak, and No. Rows in the table are sorted in decreasing order of Strong intensity to show which emotions and valences were considered most strongly related to dengue fever. The results mirror intuition, with Negative, Fear, and Anticipation representing the strongest correspondence and Anger, Surprise, and Joy representing the weakest.

Although workers were asked to rate terms on a 4-point scale, EmoLex uses a binary (0,1) score representing (Non-Evocative, Evocative). We calculate the evocation score by grouping No and Weak responses as Non-Evocative and Moderate and Strong responses as Evocative. The final two columns of Table 5 show the percentage of terms in the master set that were Non-Evocative and Evocative for each of Plutchik's eight emotions and the two valence terms.

4.2 Calculating sentiment

We use simple averaging to calculate sentiment scores for individual text blocks and aggregate text block scores over a document. It is important to subdivide the document into text blocks that are expected to contain only a single sentiment,

Table 4 Filter questions for the MTurk experiment

Filter Question 1	Filter Question 2
Which word is the closest in meaning (most related) to *mosquito*?	Based on the following description of dengue fever, which of the following answers is true? Description: Dengue is a mosquito-borne viral disease transmitted to humans through infected Aedes mosquitoes, a tropical and subtropical species that can be found throughout the world. The principal symptom of dengue is high-grade fever and can present with any of the following symptoms: facial flushing, skin erythema, body ache, myalgia, arthralgia, and severe headache. Dengue spread rapidly during the twentieth century to infect more than 300 million people in 2010. One in three people live among mosquitoes that transmit the dengue virus, yet there remain major uncertainties over the burden of dengue
Resident pest city price	Less than 10 million people were infected with dengue in 2010 1 in 3 people live among mosquitoes that transmit the dengue virus The burden of dengue is well known The principle symptom of dengue is lower back pain

Table 5 Average intensity and evocation levels for the eight emotion terms and two valence terms sorted by strong intensity

	Strong (%)	Moderate (%)	Weak (%)	No (%)	Non-evocative (%)	Evocative (%)
Negative	26.5	15.5	12.7	45.4	58.5	41.5
Fear	25.0	18.2	18.4	38.5	57.2	42.8
Anticipation	20.9	23.5	15.4	40.2	54.5	45.5
Sadness	17.5	16.7	17.2	48.7	65.6	34.4
Trust	15.2	14.2	11.3	59.0	68.5	31.5
Positive	12.4	12.9	10.6	64.1	72.7	27.3
Disgust	12.3	12.5	15.7	59.6	76.0	24.0
Anger	10.1	12.6	17.3	60.0	77.1	22.9
Surprise	5.8	9.6	15.6	69.0	85.5	14.5
Joy	5.8	6.5	10.9	76.9	87.6	12.4

for example, sentences in a newspaper article or individual tweets in a tweet set. This *averaging* ensures that opposite sentiments do not "cancel" one another, leading to neutral scores for the majority of the text blocks or documents.

Scores are calculated for each of Plutchik's eight emotions *fear*, . . ., *anticipation*. As an example, consider *joy*. The emotion score $e_{\text{joy},i,j}$ for text block $b_{i,j}$ in document d_j is calculated as follows.

$$e_{\text{joy},i,j} = \frac{\sum_{i=1}^{|b_{i,j}|} t_{i,j}(\text{joy} = 1)}{|b_{i,j}|} \tag{4}$$

where $t_{i,j}(\text{joy} = 1)$ are the terms in block $b_{i,j}$ with a *joy* score of 1 (evocative), and $|b_{i,j}|$ is the total number of terms in $b_{i,j}$. Once individual blocks are scored, they are aggregated to produce an overall *joy* score for d_j.

$$E_{\text{joy},j} = \frac{\sum_{i=1}^{|d_j|} e_{\text{joy},i,j}}{|d_j|} \tag{5}$$

where $|d_j|$ is the total number of text blocks in d_j. The same approach is used to calculate $E_{\cdot,j}$ scores for Plutchik's remaining seven emotions.

4.3 Comparison with EmoLex

Out of the 850 terms in our evaluated lexicon, 544 already exist in EmoLex, and 306 do not. The parts of speech for the new terms are (verb, 124), (noun, 102), (adjective, 71), and (adverb, 9).

For the 544 terms currently in EmoLex, Table 6 shows the number of EmoLex scores that changed for each of Plutchik's

Table 6 EmoLex scores maintained (Same) or changed (Different) following valuation, sorted by Strong intensity (Table 5)

	Same				Different			
	0	1	Total	%	0→1	1→0	Total	%
Negative	263	141	404	74.3	84	56	140	25.7
Fear	274	78	352	64.7	158	34	192	35.3
Anticipation	298	17	315	57.9	216	13	229	42.1
Sadness	335	59	394	72.4	126	24	150	27.6
Trust	366	33	399	73.3	125	20	145	26.7
Positive	378	54	432	79.4	83	29	112	20.6
Disgust	380	38	418	76.8	93	33	126	23.2
Anger	395	27	422	77.6	97	25	122	22.4
Surprise	455	7	462	84.9	67	15	82	15.1
Joy	475	4	479	88.1	57	8	65	11.9

eight emotions and the two valence terms. For example, 352 terms total remained unchanged for the Fear emotion, 158 terms switched to Evocative, and 34 terms switched to Non-Evocative. This totals to $352+158+34 = 544$, the number of evaluated terms that exist in EmoLex. The rows in Table 6 are sorted identically to Table 5: from highest Strong intensity to lowest Strong intensity. Although there is not an exact one-to-one correspondence between Strong intensity and changes in EmoLex, Table 6 shows that terms with higher Strong intensity are more likely to switch their EmoLex scores. In total, 1363 emotion values (27.1%) were reversed.

We demonstrate the emotional values for six different sentences scored with the original EmoLex dictionary and our dengue-specific EmoLex dictionary. The first two sentences show examples of unique dengue terms defining valence and emotional dimension scores. Bold blue terms exist in both dictionaries. Bold red terms exist only in the dengue-specific dictionary. The next two sentences show examples of re-evaluation of terms for a dengue context. The final two sentences show examples where EmoLex outperforms the dengue-specific dictionary. This result highlights that even a context-specific dictionary will not give the best scores in all cases.

Bold scores in Table 7 show where one dictionary contains stronger emotions versus the other for a given sentence. In general, the dengue-specific dictionary produces higher emotional scores than the original EmoLex. This finding is guaranteed to be true for sentences with only dengue-specific terms (table rows one, two, and five), but it also occurs when terms are shared between dictionaries (sentences three, four, and six). As noted above, the dengue-specific dictionary does not always produce "better" emotional scores. Consider the sixth sentence, which most readers would consider positive. In spite of this *subjective intuition*, the dengue-specific dictionary contains higher *Negative*, Fear, Anticipation, Sadness, Trust, Anger, and Surprise scores versus the original EmoLex.

A controlled comparison of the dengue-specific dictionary versus EmoLex would evaluate a subset of sentences in D and G using three or more human observers over the two valence and eight Plutchik dimensions. These form a "gold standard" baseline to allow accuracy comparisons with dengue-specific and EmoLex scores. The evaluation can be conducted using MTurk in a manner similar to how dengue unique and common terms were constructed. Unfortunately, we have not yet completed this study, so we can offer only anecdotal evidence of the value of our EmoLex extensions. The controlled evaluation is currently marked for future work.

4.4 Sentiment visualization

Our use case for the dengue-specific EmoLex dictionary was to visualize sentiment in a surveillance dashboard. Sentiment by EmoLex emotional dimension was presented using a modified Nightingale Rose chart. Rose charts, also known as Coxcomb or Polar charts, were invented by Florance Nightingale during the Crimean War to present causes of mortality [65]. The chart is circular and is subdivided into equal-angle sectors or "slices" representing a categorization of the underlying data. In Nightingale's case, the chart was divided into twelve sectors representing deaths during each month of the year. A sector's height (or radius) represents its value relative to other sectors. A sector can be further split into subcategories. For example, Nightingale's sectors represented fatalities, subdivided into wounds (red), other causes (black), and preventable (blue). Her point to the British generals was that the vast majority of fatalities were preventable through better sanitation in the field hospitals.

We visualize the progression of sentiment in the underlying data using two Rose charts: one for Twitter data and another for newspaper articles. Each Rose chart is divided into twelve sectors representing monthly data for a user-chosen year. Each sector is further divided into (up to) eight chords, representing Plutchik's eight emotional dimensions.

Table 7 Six sentences scored with the original EmoLex and the Dengue EmoLex dictionaries reporting both valence and Plutchik's emotion for *N: Negative*, F: Fear, At: Anticipation, Sd: Sadness, T: Trust, *P: Positive*, D: Disgust, Ag: Anger, Sr: Surprise, and J: Joy, bold blue words exist in both dictionaries, bold red words are unique to the Dengue dictionary, bold scores represent the larger of the two scores for a given sentence's emotions

Dict	*N*	F	At	Sd	T	*P*	D	Ag	Sr	J
*Four **dengue viral** fever **cases** were **reported** in Karachi on Friday*										
EmoLex	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Dengue	**0.3**	**0.3**	**0.1**	**0.3**	0.0	0.0	**0.2**	0.0	**0.1**	0.0
*He **informed** that **dengue** fever **occurs** soon after **rainy***										
EmoLex	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Dengue	**0.3**	**0.3**	**0.2**	**0.3**	0.0	0.0	**0.3**	**0.1**	**0.1**	0.0
*A **prevention campaign** should help with **mosquito outbreaks***										
EmoLex	0.1	0.0	0.1	0.0	0.1	0.0	0.1	0.1	0.0	0.0
Dengue	0.1	**0.3**	**0.3**	**0.3**	0.1	**0.3**	**0.3**	0.1	**0.1**	0.1
Monsoon** season and **stagnant water** are the main sources for nourishing these **diseases										
EmoLex	0.2	0.1	0.0	**0.2**	0.0	0.0	0.1	**0.1**	0.0	0.0
Dengue	**0.3**	**0.2**	**0.3**	0.1	0.0	0.0	**0.2**	0.0	0.0	0.0
***Dengue** fever **sanitation awareness** schedule carried out inadequately*										
EmoLex	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Dengue	**0.3**	**0.3**	**0.1**	**0.3**	0.0	**0.1**	**0.1**	0.0	0.0	0.0
***Admitted patients** currently in **stable medical** condition*										
EmoLex	0.0	0.1	0.3	0.0	0.1	0.3	0.0	0.0	0.0	0.0
Dengue	**0.1**	**0.4**	**0.4**	**0.3**	**0.4**	0.3	0.0	**0.1**	**0.1**	0.0

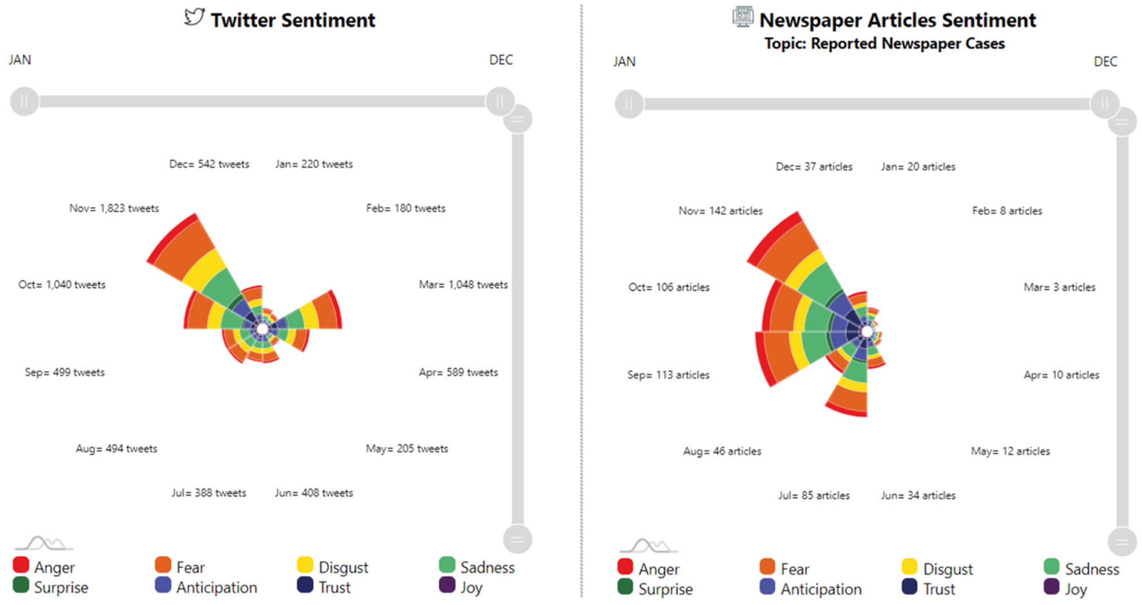

Fig. 2 Rose charts of tweet and newspaper article counts, divided by month and Plutchik's emotional dimensions

The height of a chord represents the percentage of the total sentiment score collected for the given month. The height of a sector represents the percentage of the total newspaper articles collected for the given year.

Figure 2 shows data for 2014, using a tweet dataset and a newspaper article dataset. Scroll bars on the top and right of each visualization allow filtering the months being visualized and zooming the visualization for better analysis of small chords. Hovering over a chord provides a tooltip with information about the number and percentage of documents categorized into the particular emotion within the month. The visualization provides some interesting insights into estimated dengue counts. For example, November has the largest counts in both tweets and newspaper articles; the Fear emotion is most prominent across most months; very little newspaper activity occurs in the January to June timeframe, but tweets spike in March and April; the number of tweets are approximately an order of magnitude higher than the number of newspaper articles; the relative pattern of emotions in each sector is relatively similar between tweets and newspaper articles, although not identical. Follow-on analysis based on these findings has the potential to uncover useful conclusions about how different communication channels (e.g., social media versus traditional media) are used in different ways.

5 Discussion

This paper presents a semiautomated method to extend a general sentiment dictionary for a user-chosen domain of interest. We describe how to identify both unique terms to add to the dictionary and existing terms that may require re-evaluation of their sentiment scores.

Term sentiment scores for Plutchik's emotional dimensions are obtained using Amazon Mechanical Turk. We demonstrate our sentiment scores assigned to estimated dengue cases in India, visualized using Rose diagrams. The results were positive, both in terms of the small number of terms identified for evaluation and addition and the improvements the extensions provided versus the original EmoLex dictionary.

6 Conclusions & Future work

We demonstrate how the results of an extended dictionary can be further used for estimating case count patterns in two domains: dengue fever in India and influenza in the USA. The positive results would suggest that the extended dictionary offers accuracy and specificity not available in its general form. We also highlight existing areas of weakness in our current algorithm, together with suggestions for how these might be addressed in future work.

6.1 Estimating dengue case counts

Although the focus of this paper is not on how dengue case counts are predicted, we provide a brief overview of how our sentiment estimates are being integrated into a dengue surveillance system [66]. The surveillance system uses English language newspaper articles to track dengue outbreaks in an accurate and timely fashion. This type of surveillance is especially important in countries that report little or no information about the current dengue status. Our basic monthly dengue topic identification strategy to estimate dengue case counts is described below.

1. Brandwatch English language Indian newspaper articles on dengue are collected and divided by month.
2. For each month m, Latent Dirichlet Analysis with Differential Evolution (LDADE) is applied to determine an optimal topic count k_m and priors for topic and word distribution α_m and β_m [67].
3. LDA is used to extract k_m topics for each month. Collaboration with a domain expert is used to identify keywords to label each topic (Table 8) [68].
4. Extracted topics are used to build "topic graphs" of the monthly frequency of topics in historical dengue newspaper articles over five regions in India: north, south, east, west, and central.
5. Different algorithms (Pearson, Spearman, Kendall Tau-b, and Maximal Information Coefficient or MIC) showed a strong positive correlation between newspaper case topic counts and reported dengue cases. We substitute topic counts as a proxy for dengue cases.
6. To estimate dengue cases for unreported years, newspaper articles from these years are topic classified to estimate topic counts, which are converted to estimated dengue case counts. Analysis of supervised machine learning algorithms identified Naïve Bayes classification using bag-of-words feature extraction as most accurate.

Since official dengue counts do not exist, there was no way to validate the accuracy of our results. To address this *issue*, we obtain flu rates from the CDC's Outpatient Influenza-Like Illness Surveillance Network (ILINet), together with newspaper articles discussing influenza in the United States. Our predicted flu case counts with a one-month lag showed correlation rates to known influenza rates of 90%, 86%, 71%, and 98% for Pearson, Spearman, Kendall Tau-b, and MIC, respectively.

The sentiment Rose charts are one part of the overall surveillance system. Other visualizations include streamgraphs to show topic volume over time and to compare against known dengue indicators (e.g., precipitation). Line graphs allow for comparing the monthly patterns of Plutchik's emotional dimensions. We also investigated how large the document set needs to be to produce acceptable performance. Our influenza dataset initially contained 135,658 articles, retrieved with queries to the Brandwatch database using keywords suggested by influenza experts we collaborated with during the project. The same keywords were used to retrieve historical Twitter data over an identical time period using Brandwatch[3] (formerly Crimson Hexagon), a subscription service that provides "insights from 100 million sources and 1.4 trillion posts." Analysis of downsizing the collection to simulate fewer documents suggested a reduction of up to 50% can still produce the acceptable results. Below that,

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

Table 8 Keywords and labels for $k_{\text{Jan}} = 3$, $k_{\text{Apr}} = 3$, $k_{\text{Jul}} = 2$, and $k_{\text{Oct}} = 2$ topics from Indian dengue newspaper articles

Month	Topic Keywords	Topic Label
January	Area, government, new, people, waste, water, year	Politics
	Area, doctor, fogging, hospital, mosquito, officer, resident	Prevention
	Case, disease, healthy, malaria, number, reported, year	Reported cases
April	Blood, government, healthy, hospital, medical, medicine, patient	Politics
	Area, case, city, corporation, mosquito, water, year	Prevention
	Disease, case, health, malaria, mosquito, vectorborne, year	Reported cases
July	Case, fever, healthy, hospital, mosquito, state, year	Prevention
	Case, city, disease, healthy, malaria, mosquito, water	Reported cases
October	Area, city, department, health, hospital, mosquito, water	Prevention
	Case, disease, health, hospital, number, patient, test	Reported cases

performance begins to degrade more rapidly. Finally, evaluation of the surveillance system using the influenza dataset and influenza experts produced strong support for both the system's capabilities and the trends and patterns it produced [66,69].

In addition to improving the capabilities of a sentiment dictionary, extending sentiment to a specific domain can have important advantages for research in that domain. For example, surveillance systems for epidemiological diseases like dengue are critical since timely identification of the onset and spread of the disease is often not available. Through the use of public media and social network sources, current information can be provided to the general public.

6.2 Limitations

Although results-to-date are promising, several limitations exist in our current approach. We are now investigating ways to address these issues as an area of future work.

1. *Automation* Although we worked to minimize manual effort, a small number of terms must be evaluated in the context of the target domain to update or extend a general sentiment dictionary. We are pursuing a strategy to examine a broad collection of sentiment dictionaries for terms to be added to EmoLex. This approach introduces the problem of converting different types of sentiment estimates to Plutchik's emotional dimensions. It also does not address the need to re-evaluate existing terms in the context of a target domain.
2. *Sentiment Dictionaries versus Human Evaluation* We are designing a controlled experiment in MTurk to evaluate the accuracy of our dictionary scores versus human scoring since human evaluation is often considered the "gold standard" for accuracy calculations and identification of estimation errors.
3. *Social Media Text* EmoLex was not designed to evaluate social media text. Since we intend to use social media input in our surveillance system, we need to generate sentiment estimates for common social media elements like emoticons and social media abbreviations.
4. *Dynamic Updates* Currently, our approach does not include real-time data injection. Updating the dataset dynamically is easy to do, but evaluating new unique and common terms requires time to run an MTurk experiment. Fortunately, we can bootstrap the process by ignoring terms that have already been evaluated. We suspect this process will significantly reduce the number of new terms to only a few, allowing us to perform evaluation on a less frequent schedule.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Alharbi, M., Laramee, R.S.: SoS TextViz: an extend survey of surveys on text visualization. Computers **8**(1), 143–152 (2019)
2. Dou, W., Liu, S.: Topic- and time-oriented visual text analysis. IEEE Comput. Gr. Vis. **36**(4), 8–13 (2016)
3. Kucher, K., Paradis, C., Kerren, A.: State of the art in sentiment visualization. Comput. Gr. Forum **37**(1), 71–96 (2017)
4. Shepard, D.S., Halasa, Y.A., Tyagi, B.K., Adhish, S.V., Nandan, D., Karthiga, K.S., Chellaswamy, V., Gaba, M., Arora, N.K.: Economic and disease burden of dengue illness in India. Am. J. Trop. Med. Hyg. **91**(6), 1235–1242 (2014)
5. Plutchik, R.: A general psychoevolutionary theory of emotion. In: Plutchik, R., Kellerman, H. (eds.) Theories of Emotion?: Emotion, Theory, Research, and Experience, pp. 3–31. Academic Press, New York (1980)
6. Mohammad, S.M., Turney, P.D.: Crowdsourcing a word-emotion association lexicon. Comput. Intell. **29**(3), 436–465 (2013)
7. Liu, B., Zhang, L.: A survey of opinion mining and sentiment analysis. In: Aggarwal, C.C., Zhai, C.X. (eds.) Mining Text Data, pp. 415–463. Springer, New York (2012)

8. Mohammad, S.M.: Sentiment analysis: detecting valence, emotions, and other affectual states from text. In: Meiselman, H. (ed.) Emotional Measurement, pp. 201–237. Elsevier, Atlanta (2015)
9. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Found. Trends Inf. Retr. **2**(1–2), 1–135 (2008)
10. Zhang, L., Wang, S., Liu, B.: Deep learning for sentiment analysis: a survey. WIREs Data Min. Knowl. Discov. **8**(4), 1–25 (2018)
11. Russell, J.A.: A circumplex model of affect. J. Personal. Soc. Psychol. **39**(6), 1161–1178 (1980)
12. Russell, J.A., Feldman Barrett, L.: The structure of current affect: controversies and emerging consensus. Curr. Dir. Psychol. Sci. **8**(1), 10–14 (1999)
13. Pang, B., Lee, L.: A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of the 42nd annual meeting of the association for computational linguistics (ACL '04), Barcelona, Spain, pp. 271–278 (2004)
14. Pang, B., Lee, L., Vithyanathan, S.: Thumbs up? Sentiment classification using machine learning techniques. In: Proceedings of the 2002 conference on empirical methods in natural language processing (EMNLP 2002), Philadelphia, PA, pp. 79–86 (2002)
15. Turney, P.: Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the 40th annual meeting of the association for computational linguistics (ACL '02), Philadelphia, PA, pp. 417–424 (2002)
16. Bonata, V., Janardhan, N.: A comprehensive study on lxicon based approaches for sentiment analysis. Asian J. Comput. Sci. Technol. **8**(S2), 1–6 (2019)
17. DiBattista, J.: The best python sentiment analysis package (+1 Huge Mistake). https://towardsdatascience.com/the-best-python-sentiment-analysis-package-1-huge-common-mistake-d6da9ad6cdeb. Online; accessed 02 Mar 2021 (2021)
18. Podiotis, P.: Sentiment analysis of the CIA world Factbook). Social science research network (SSRN), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3721400. Online; accessed 02 Mar 2021 (2020)
19. Li, Z., Wei, Y., Zhang, Y., Yang, Q.: Hierarchical attention transfer network for cross-domain sentiment classification. In: Proceedings of the thirty-second AAAI conference on artifical intelligence (AAAI-18), New Orleans, LA, pp. 5852–5859 (2018)
20. Zhang, K., Zhang, K., Zhang, M., Zhao, H., Liu, W., Wei, W.: Incorporating dynamic semantics into pre-trained language model for aspect-based sentiment analysis. In: Cohn, T., He, Y., Liu, Y. (eds.) Findings of the Association for Computational Linguistics (ACL 2022), pp. 3599–3610. Ireland, Dublin (2022)
21. Kenton, J.D., Chang, M.-W., Toutanova, L.K.: BERT: Pre-training of deep bidirectional transforms for language understanding. In: Proceedings of the 2019 annual conference of the North American chapter of the association for computational linguistics-human language technologies (NAACL-HLT 2019), virtual, pp. 4171–4189 (2019)
22. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners. In: Advances in Neural Information Processing Systems 33 (NeurIPS 2020), pp. 1877–1901. virtual, (2020)
23. Lewis, M., Liu, Y., Goya, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Proceedings of the 58th annual meeting of the association for computational linguistics (ACL 2020), Seattle, Washington, pp. 7871–7880 (2020)
24. Song, K., Tan, X., Qin, T., Lu, U., Y., L.T.: MASS: Masked sequence to sequence pre-training for language generation. In: Proceedings of the 36th international conference on machine learning (ICML 2019), Long Beach, California, pp. 5926–5936 (2019)
25. Pepe, A., Bollen, J.: Between conjecture and memento: shaping a collective emotional perception of the future. In: AAAI spring symposium on emotion, personality, and social behavior, Stanford, CA, pp. 111–116 (2008)
26. Mislove, A., Lehmann, S., Ahn, Y.-Y., Onnela, J.-P., Rosenquist, J.N.: Pulse of the Nation: U.S. Mood Throughout the Day Inferred from Twitter. http://www.ccs.neu.edu/home/amislove/twittermood (2010)
27. Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., Kappas, A.: Sentiment strength detection in short informal text. J. Am. Soc. Inf. Sci. Technol. **61**(12), 2544–2558 (2010)
28. Tausczik, Y.R., Pennebaker, J.W.: The psychological meaning of words: LIWC and computerized text analysis methods. J. Lang. Soc. Psychol. **29**(1), 25–54 (2010)
29. Baccianella, S., Esuli, A., Sebastiani, F.: SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In: Proceedings of the 7th international conference on language resources and evaluation (LREC '10), Valletta, Malta, pp. 2200–2204 (2010)
30. Warriner, A.B., Kuperman, V., Brysbaert, M.: Norms of valence, arousal, and dominance for 13,915 English lemmas. Behav. Res. Methods **45**(4), 1191–1207 (2013)
31. Cao, N., Lin, Y.-R., Sun, X., Lazer, D., Liu, S., Huamin, Q.: Whisper: Tracing the spatiotemporal process of information diffusion in real time. IEEE Trans Vis. Comput. Gr. **18**(12), 2649–2658 (2012)
32. Cao, N., Lu, L., Lin, Y.-R., Wang, F.: SocialHelix: Visual analysis of sentiment divergence in social media. J. Vis. **18**(2), 221–235 (2014)
33. Wu, Y., Liu, S., Yan, K., Liu, M., Wu, F.: OpinionFlow: visual analysis of opinion diffusion on social media. IEEE Trans. Vis. Comput. Gr. **20**(12), 1763–1772 (2014)
34. Liu, Y., Wang, H., Landis, S., Macjejewski, R.: A visual analytics framework for identifying topic drivers in media events. IEEE Trans. Vis. Comput. Gr. **24**(9), 2501–2515 (2017)
35. El-Assady, M., Gold, V., Acevedo, C., Collins, C., Keim, D.: ConToVi: multi-party conversation exploration using topic-space views. Comput. Gr. Forum **35**(3), 431–440 (2016)
36. El-Assady, M., Sevastjanova, R., Keim, D., Collins, C.: Thread-Reconstructor: modeling reply-chains to untangle conversational text through visual analytics. Comput. Gr. Forum **37**(3), 351–365 (2018)
37. Hoque, E., Carenini, G.: ConVis: a visual text analytic system for exploring blog conversations. Comput. Gr. Forum **33**(3), 221–230 (2014)
38. Hoque, E., Carenini, G.: MultiConVis: A visual text analysis system for exploring a collection of online conversations. In: Proceedings of the 21st international conference on intelligent user interfaces (IUI '16), Sonoma, CA, pp. 96–107 (2016)
39. Mohammad, S.M., Sobhani, P., Kiritchenko, S.: Stance and sentiment in tweets. ACM Trans. Int. Technol. **17**(3), 26 (2017)
40. Kucher, K., Martins, R.M., Paradis, C., Kerren, A.: StanceVis Prime: visual analysis of sentiment and stance in social media texts. J. Vis. **23**(6), 1015–1034 (2020)
41. Wei, F., Shixia, L., Yangqiu, S., Shimei, P., Zhou, M.X., Qian, W., Lei, S., Li, T., Qiang, Z.: TIARA: interactive, topic-based visual text summarization and analysis. In: Proceedings of the 16th SIGKDD international conference on knowledge discovery and data mining (KDD 2010), Washington, DC, pp. 153–162 (2010)
42. Dörk, M., Gruen, D., Williamson, C., Carpendale, S.: A visual backchannel for large-scale events. IEEE Trans. Vis. Comput. Gr. **16**(6), 1129–1138 (2010)

43. Mohammad, S.M.: Challenges in sentiment analysis. In: Das, D., Cambria, E., Bandyopadhyay, S. (eds.) A Practical Guide to Sentiment Analysis, pp. 61–83. Springer, New York (2016)
44. World Health Organization: Prevention and control of dengue and dengue hemorrhagic fever: comprehensive guidelines. Technical report, World Health Organization Regional Office for South-East Asia (1999)
45. Bhatt, S., Gething, P.W., Brady, O.J., Messina, J.P., Farlow, A.W., Moyes, C.L., Drake, J.M., Brownstein, J.S., Hoen, A.G., Sankoh, O.: The global distribution and burden of dengue. Nature **496**(7446), 504 (2013)
46. Montoya, M., Gresh, L., Mercado, J.C., Williams, K.L., Vargas, M.J., Gutierrez, G., Kuan, G., Gordon, A., Balmaseda, A., Harris, E.: Symptomatic versus inapparent outcome in repeat dengue virus infections is influenced by the time interval between infections and study year. PLoS Negl. Trop. Dis. **7**(8), 2357 (2013)
47. Moreira, L.A., Iturbe-Ormaetxe, I., Jeffery, J.A., Lu, G., Pyke, A.T., Hedges, L.M., Rocha, B.C., Hall-Mendelin, S., Day, A., Riegler, M.: A Wolbachia symbiont in Aedes Aegypti limits infection with dengue, chikungunya, and plasmodium. Cell **139**(7), 1268–1278 (2009)
48. Olkowski, S., Forshey, B.M., Morrison, A.C., Rocha, C., Vilcarromero, S., Halsey, E.S., Kochel, T.J., Scott, T.W., Stoddard, S.T.: Reduced risk of disease during postsecondary dengue virus infections. J. Infect. Dis. **208**(6), 1026–1033 (2013)
49. Reyes, M., Mercado, J.C., Standish, K., Matute, J.C., Ortega, O., Moraga, B., Avilés, W., Henn, M.R., Balmaseda, A., Kuan, G.: Index cluster study of dengue virus infection in Nicaragua. Am. J. Trop. Med. Hyg. **83**(3), 683–689 (2010)
50. Shepard, D.S., Undurraga, E.A., Halasa, Y.A.: Economic and disease burden of dengue in southeast asia. PLoS Negl. Trop. Dis. **7**(2), 2055 (2013)
51. Lozano, R., Naghavi, M., Foreman, K., Lim, S., Shibuya, K., Aboyans, V., Abraham, J., Adair, T., Aggarwal, R., Ahn, S.Y.: Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. Lancet **380**(9859), 2095–2128 (2012)
52. World Health Organization: Setting priorities in communicable disease surveillance. Technical report, World Health Organization, Lyon, France (2006)
53. Brownstein, J.S., Freifeld, C.C., Reis, B.Y., Mandl, K.D.: Surveillance sans frontières: internet-based emerging infectious disease intelligence and the HealthMap project. PLoS Med. **5**(7), 151 (2008)
54. Davies, S.E.: The challenge to know and control: disease outbreak surveillance and alerts in China and India. Glob. Pub. Health **7**(7), 695–716 (2012)
55. Farrington, C.P., Andrews, N.J., Beale, A.D., Catchpole, M.A.: A statistical algorithm for the early detection of outbreaks of infectious disease. J. Royal Stat. Soc. Series A (Statistics in Society) **159**(3), 547–563 (1996)
56. Liu, Y.: China's public health-care system: facing the challenges. Bull. World Health Organ. **82**(7), 532–538 (2004)
57. Thacker, S.B., Qualters, J.R., Lee, L.M.: Public health surveillance in the United States: evolution and challenges. MMWR Surveill. Summ. **61**, 3–9 (2012)
58. Beatty, M.E., Stone, A., Fitzsimons, D.W., Hanna, J.N., Lam, S.K., Vong, S., Guzman, M.G., Mendez-Galvan, J.F., Halstead, S.B., Letson, G.W.: Best practices in dengue surveillance: a report from the Asia-Pacific and Americas dengue prevention boards. PLoS Negl. Trop. Dis. **4**(11), 890 (2010)
59. Konowitz, P.M., Petrossian, G.A., Rose, D.N.: The underreporting of disease and physicians' knowledge of reporting requirements. Pub. Health Rep. **99**(1), 31 (1984)
60. McKenzie, J.F., Pinger, R.R.: An Introduction to Community Health, Brief Jones & Bartlett Publishers, Burlington (2013)
61. Singh, J., Dinkar, A., Atam, V., Himanshu, D., Gupta, K.K., Usman, K., Misra, R.: Awareness and outcome of changing trends in clinical profile of dengue fever: a retrospective analysis of dengue epidemic from January to December 2014 at a tertiary care hospital. J. Assoc. Phys. India **65**, 42 (2017)
62. Fisher, R.A.: Statistical Methods for Research Workers. Oliver & Boyd, Edinburugh (1925)
63. Upton, G.J.: Fisher's exact test. J. Royal Stat. Soc. Series A **155**(3), 395–402 (1992)
64. Kelly, J.T., Loepp, E.: Distinction without a difference? An assessment of MTurk worker types. Res. Polit. (2020). https://doi.org/10.11772/2053168019901185
65. Sherlock, A.: Florence Nightingale's "Rose" Diagram (2021). https://www.maharam.com/stories/sherlock_florence-nightingales-rose-diagram
66. Villanes, A., Griffiths, E., Rappa, M., Healey, C.G.: Dengue fever surveillance in India using text mining in public media. Am. J. Trop. Med. Hyg. **98**, 181–191 (2018)
67. Agarwal, A., Fu, W., Menzies, T.: What is wrong with topic modeling? And how to fix it using search-based software engineering. Inf. Softw. Technol. **98**, 74–88 (2018)
68. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. **3**(4–5), 993–1022 (2003)
69. Villanes, A.: Epidemiological disease surveillance using public media text mining. PhD thesis, North Carolina State University (2019)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.