



City Research Online

City St George's, University of London

Citation: K., Anoop, P., Deepak, S., Sam Abraham, V. L., Lajish & M. P., Gangan (2022). Readers' affect: predicting and understanding readers' emotions with deep learning. *Journal of Big Data*, 9(1), 82. doi: 10.1186/s40537-022-00614-2

This is the published version of the paper.

This version of the publication may differ from the final published version. To cite this item please consult the publisher's version.

Permanent repository link: <https://openaccess.city.ac.uk/id/eprint/36025/>

Link to published version: <https://doi.org/10.1186/s40537-022-00614-2>






Copyright and Reuse: Copyright and Moral Rights remain with the author(s) and/or copyright holders. Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge, unless otherwise indicated, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way. For full details of reuse please refer to [City Research Online policy](#).

RESEARCH

Open Access



Readers' affect: predicting and understanding readers' emotions with deep learning

Anoop K.¹ , Deepak P.^{2*} , Savitha Sam Abraham³ , Lajish V. L.¹  and Manjary P. Gangan¹ 

*Correspondence:
deepaksp@acm.org

¹ Department of Computer Science, University of Calicut, Malappuram, Kerala, India

² School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, Belfast, Northern Ireland, UK

³ School of Science and Technology, Örebro University, Örebro, Sweden

Abstract

Emotions are highly useful to model human behavior being at the core of what makes us human. Today, people abundantly express and share emotions through social media. Technological advancements in such platforms enable sharing opinions or expressing any specific emotions towards what others have shared, mainly in the form of textual data. This entails an interesting arena for analysis; as to whether there is a disconnect between the writer's intended emotion and the reader's perception of textual content. In this paper, we present experiments for Readers' Emotion Detection through multi-target regression settings by exploring a Bi-LSTM-based Attention model, where our major intention is to analyze the interpretability and effectiveness of the deep learning model for the task. To conduct experiments, we procure two extensive datasets REN-10k and RENh-4k, apart from using a popular benchmark dataset from SemEval-2007. We perform a two-phase experimental evaluation, first being various coarse-grained and fine-grained evaluations of our *model performance* in comparison with several baselines belonging to different categories of emotion detection, viz., deep learning, lexicon based, and classical machine learning. Secondly, we evaluate *model behavior* towards readers' emotion detection assessing attention maps generated by the model through devising a novel set of qualitative and quantitative metrics. The first phase of experiments shows that our Bi-LSTM + Attention model significantly outperforms all baselines. The second analysis reveals that emotions may be correlated to specific words as well as named entities.

Keywords: Readers' emotion detection, Affective computing, Textual emotion detection, Deep learning, Attention, Interpretability

Introduction

The rise of social media and advancements in information technology enables millions of individuals to write, share, or even criticize opinions freely. This produces a deluge of social interactions manifested through textual data. The ability to add expressive opinions scattered with emojis makes it easy to express diverse emotions easily. The

expression of emotions on social media has been modulated by new affordances from social media platforms such as when Facebook in 2016 introduced five main emotion reactions to deepen embedding of emotions in responses to social media posts¹. The presence and usage of such affordances provide a wealth of data to analyze and offers space for research into textual data through different perspectives, such as the *emotion expressed* by the writer (Writer Emotion), the *emotion elicited* from the readers' (Readers' Emotion), and the *dichotomy between expressed and perceived emotions in textual emotion detection*. This is because in most cases readers' emotions triggered by the document do not always agree with the writer emotions. Leveraging readers' emotions has numerous potential applications that have attracted attention from the Natural Language Processing (NLP) and machine learning research sub-communities through a variety of tasks, viz., emotion aware search engines/recommendation systems, emotion enriched article generation, automated article editing to filter out or diminish the emotionally sensitive contents, forecasting readers' emotions on any creative article so that the writer can realize emotions that influence the readers' in advance, etc., [1, 2].

The computational task of detecting readers' emotions is generally formulated as a single/multi-class or multi-label classification task [3–7]. A minority of approaches model the task as a multi-target method [8–10], that usually follows the traditional NLP regression settings and helps to gain information on the intensity of corresponding emotions, apart from detecting emotion classes. Research into readers' emotion detection take advantage of methods such as lexicon based [11], rule-based decision making [3], classical machine learning [12, 13], deep learning [10, 14], and hybrid approaches [15]. Of these, deep learning based approaches are usually observed to outperform other approaches, as generally in the case of many other areas of NLP including text classification, machine translation, sentiment analysis, etc., [16–18], with the advent of various architectures such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) like Long Short-Term Memory (LSTM) and Bidirectional Long Short-Term Memory (Bi-LSTM), that encompass multiple levels of non-linear operations to accommodate automated feature representation of input data with different hierarchies of abstraction. Among the deep learning based studies in textual emotion detection, there has been some recent interest in utilizing attention mechanisms to improve model performance [19] or to observe the words responsible for decision making [20]. But, to our best knowledge, there has been no prior work analyzing and quantifying the role of emotion words or named entities for the task of readers' emotion detection. In this work, we utilize a Bi-LSTM + Attention model with an intention to analyze the interpretable nature and behavior of the model for readers' emotion detection through multi-target regression settings over short-text news documents, where we perform detailed qualitative and quantitative analysis to understand the underlying model behavior and to quantify the role of emotion words and named entities in decision making. The major benefits of our study include a readers' emotion detection model that performs better than the baselines, systematic investigation of the model's decision making (model behavior) and specifically studying the role of emotion words and named entities for the

¹ <https://about.fb.com/news/2016/02/reactions-now-available-globally/>.

task. In this study, to represent readers' emotions we utilize the discrete basic emotions defined by Paul Ekman [21] (*happiness, sadness, anger, fear, disgust, and surprise*), since they are the most frequently discussed basic emotions by the theorists in discrete emotion models, and also, most of the social media platforms allow their users to react to news or posts with discrete emotion representations.

Motivation and contributions

Inspired by recent works in the related area of sentiment analysis proposed by Kardakis et al. [19] to investigate the performance improvement of the attention based deep neural networks over non-attention based models, and the work by Sen et al. [22] to explore the interpretability of attention based deep neural networks, our objective in this work specific to the task of readers' emotion detection is to evaluate the attention enabled deep neural architecture and to illustrate that attention models have the potential to enrich the model prediction while enhancing the understanding of the process of decision making. Hence, in this work we limit the investigations towards readers' emotion detection using attention enabled Bi-LSTM. Other state-of-the-art technologies such as transformer-based language models are outside the scope of our present study.

To our best knowledge, there are only a few datasets that provide emotion intensities for regression based studies [23, 24]. However, these datasets are not suitable for multi-target regression settings specific to readers' emotion detection as they map documents to only a single emotion with corresponding intensity. An available benchmark dataset that suits multi-target regression based readers' emotion detection is the SemEval-2007 [25], but being annotated by only six readers, this dataset doesn't meet the real-world scenario of a document being read and annotated by many readers. Also, even though there are few readers' emotion detection models that have been benchmarked over specific languages (e.g., [13, 26] that utilize Chinese corpora), there exists a need for readers' emotion detection dataset in English to learn the linguistic and affective characteristics within English text. This inadequacy, as also mentioned in [12, 27, 28], motivates us to procure extensive datasets that particularly suit the deep learning based multi-target regression settings to predict readers' emotion intensities rather than emotion class mapping.

The major contributions of this work are:

- We explore a Bi-LSTM + Attention model for the task of readers' emotion detection through multi-target regression settings over short-text news documents and compare the model performance against a set of baselines belonging to various families of textual emotion detection techniques including lexicon based, machine learning, and deep learning, using an extensive set of *coarse-grained* and *fine-grained evaluation measures*
- We investigate interpretability of the attention mechanism to understand the underlying behavior of Bi-LSTM + Attention model for the task of readers' emotion detection by conducting *qualitative and quantitative analysis* to quantify the role of emotion words and named entities in the model's decision making.
- We procure two new readers' emotion news datasets, REN-10k and RENh-4k where the news articles are associated with corresponding readers' emotions. We

also assign the associated genre information to the articles. As a result, apart from readers' emotion detection, these datasets can be used for multiple tasks including, document summarization and genre classification, in various scales (short-text and long-text), making them *heterogeneous task datasets*. We shall contribute REN-10k at <https://dcs.uoc.ac.in/cida/resources/ren-10k.html> and RENh-4k at <https://dcs.uoc.ac.in/cida/resources/renh-4k.html> publicly, along with the publication to aid future research.

The rest of the paper is organized as, the review of literature presented in “[Related work](#)” section, followed by methodology in “[Multi-target readers' emotion detection](#)” section, dataset description, experimental setup, model performance evaluation, and model behavior analysis in “[Empirical study](#)” section, and finally, the concluding remarks with scope for future research in “[Conclusion](#)” section.

Related work

Among the large volume of studies present in literature for textual emotion detection, including the writer/document perspective and readers' perspective, only a few focus on readers' perspective of textual emotion detection. In this section, we review prominent works in the writer and readers' perspective of textual emotion detection across three categories, viz., lexicon based, classical machine learning, and deep learning approaches. The abundance of work using deep learning prompts us to consider it as a separate category despite it falling within the broader machine learning umbrella.

Lexicon based approaches

Studies in this context leverage emotion lexicons, including general-purpose [29–31] and domain-specific emotion lexicons [32], which consist of lexical word units and their intensity associations to the emotion classes, and its utility to build numerous emotion detection systems by exploiting word level matches. There has been limited exploration in the lexicon based approach of textual emotion detection, very specific to readers' emotions. Such readers' emotion detection works began with the popular shared task, SemEval-2007 Task 14 [25], to predict the intensity of different emotion classes for a reader annotated dataset, where SWAT [11] is one of the popular among the top three systems of this task. This was followed by other works like the Emotion–Term model built over Naïve Bayes and its extension, the Emotion-Topic model that uses topic models [12]. Even though lexicon based approaches are beneficial enough due to their simplicity and ease of spotting keywords from the relevant vocabulary, they are limited in their ability towards handling negations, multiple word senses etc. In this context, Krcadinac et al. [33] illustrates the possibility of a hybrid lexicon based system, *Synesketch*, with several heuristic rule sets along with emotion lexicons for textual emotion detection, even though not specifically for readers' emotion. We make use of *Synesketch* [33], and two other promising lexicon based approaches specific to readers' emotion detection, i.e., SWAT [11] and Emotion-Term Model [12], as baselines for model performance comparison.

Machine learning based approaches

Classical machine learning opens up the way to learn hidden patterns in data through several mathematical models and overcome the drawbacks of lexicon based approaches in handling words with implicit emotion expressions. Most studies in this approach of textual emotion detection are designed as supervised multi-class tasks and some as multi-label/target tasks [7], with learning models like Support Vector Machine (SVM) [34], Naïve Bayes [35], multi-layer perceptron [36], logistic regression [37, 38] etc. Features used across such approaches can be broadly categorized as Linguistic features [34, 39], Symbol level features [32], and Affective features [32, 40]. Apart from widely explored linguistic features like TF-IDF, N-grams, BOW, etc., Ren et al. [39] utilizes pre-trained word embeddings for computing Word Mover's Distance (WMD), a distance based feature to address textual emotion detection. Readers' perspective of textual emotion detection also rely on almost the same set of features and learning prototypes for multi-class [1, 2] and multi-label/target [4, 5] settings. Apart from the supervised studies, there also exists unsupervised ways of readers' emotion detection built with the help of topic level parameters [12, 13, 27]. But Dong et al., points out that such topic-level works are more suitable to predict writer emotion rather than readers' emotions [14]. Considering these, we choose baseline models that follow multi-target regression based settings since those are likely more suitable to predict readers' emotion intensities, rather than simply mapping to the emotion classes as done in multi-class/label classification settings. Multi-target problems can be addressed in many ways like problem transformation, algorithm adaptation, and ensemble approaches [41]; we use baselines that leverage both problem transformation and algorithm adaptation with a few prominent linguistic and affective features.

Deep learning based approaches

Deep learning architectures significantly outperform classical machine learning methods in most NLP tasks off late. Deep learning based works in textual emotion detection includes CNNs [42], combination of CNN with various RNN models [15, 43], stacked RNNs [44], attention-based architectures [45], Gated Recurrent Unit (GRU) [46], LSTM [47], etc. Apart from these studies, Kratzwald et al. [48] and Chatterjee et al. [44] consider the possibilities of sentiment aided transfer learning (sent2affect) and sentiment-specific word embedding (SS-BED), respectively, for textual emotion detection. Research in textual emotion detection specific to readers' emotions also explore similar learning architectures [9, 14, 49]. Slightly different lines of inquiry to predict readers' emotions are presented in recent works, viz., [50] that utilize an ontology driven knowledge base with deep learning classifier and [51] that combines comments along with articles as input to their deep learning model. In reference to such recent advances, we draw upon the notable studies sent2affect [48] and SS-BED [44], and the RNN architectures, GRU [46], LSTM and Bi-LSTM [15, 44, 48], as baselines in our empirical evaluation.

The question of interpretability

Deep learning based approaches for textual emotion detection are found to generally outperform other approaches but, their decisions are not easily explainable as their core

learnings are embedded deep within several weight parameters. Nonetheless, there has been much interest in using attention networks in order to throw light into the workings of deep learning models. Using attention, neural architectures can automatically differentiate slices of input data in form of weights, and such learnt attention can also aid the overall learning. This helps to boost overall model performance and enhance interpretability. While there has been research in textual emotion detection that incorporate attention mechanisms to improve model performance [43, 52] or to observe salient words responsible for decision making in typical architectures [45, 53, 54], there has been virtually no exploration tuned specifically to readers' emotion detection; however, models for related tasks may be considered for the task. The sentiment analysis based work by Sen et al. [22] demonstrating and quantifying the resemblance of machine attention maps with hand-labeled human attention maps is a notable work in this regard. Others include research on text classification by Lertvittayakumjorn et al. [55] that performs human grounded explanation evaluations to analyze model behavior, model predictions, and uncertain predictions, and the research by Wiegrefe et al. [56] proposing various tests to determine the usefulness of attention to obtain explanations. Insights from these works along with some of the attention based works in NLP (e.g., [19, 57]) show that attention does encode several linguistic notions and hence one can utilize attention as a prominent way of interpretability to open the neural black box. In this context, our study adopts an attention mechanism for readers' emotion detection to interpret emotion associated linguistic notions and their importance in predictions.

Multi-target readers' emotion detection

We now outline our task more formally. We formulate the task of detecting readers' emotions of a textual document as a multi-target regression problem, where the statistical model applied on each input document is expected to produce intensity values for various emotion classes namely, *anger*, *fear*, *joy*, *sadness*, and *surprise*. Each textual document d consists of a sequence of words $[w_1, w_2, w_3, \dots]$, each word drawn from the dictionary of words compiled from across the document corpus.

For each d , the corresponding readers' emotion profile from labelled data is modelled as a normalized distribution of votes cast by multiple readers' for E distinct emotions represented as,

$$ep_r(d) = \{e_1, e_2, \dots, e_E\} \in \mathbb{R}^E \text{ where } e_i \in [0, 1] \text{ and } \sum_i e_i = 1 \quad (1)$$

Thus, a document that has gathered equal votes for a set of five emotions would yield $ep_r(d) = [0.2, 0.2, 0.2, 0.2, 0.2]$. The sum-to-one normalization enables placing documents of different popularity (i.e., vote abundance) on the same footing. Thus, the labelled corpus D with M documents can be represented as, $D = \{(d_1, ep_r(d_1)), (d_2, ep_r(d_2)), \dots, (d_M, ep_r(d_M))\}$, where, $ep_r(d_i)$ indicates the readers' emotion profile of document d_i .

The supervised task of reader-emotion detection is then to find the best fit mapping function $f : \text{document} \rightarrow \mathbb{R}^E$, such that each document d is mapped as close as possible to the readers' emotion profile from the labelled data, i.e., $ep_r(d)$.

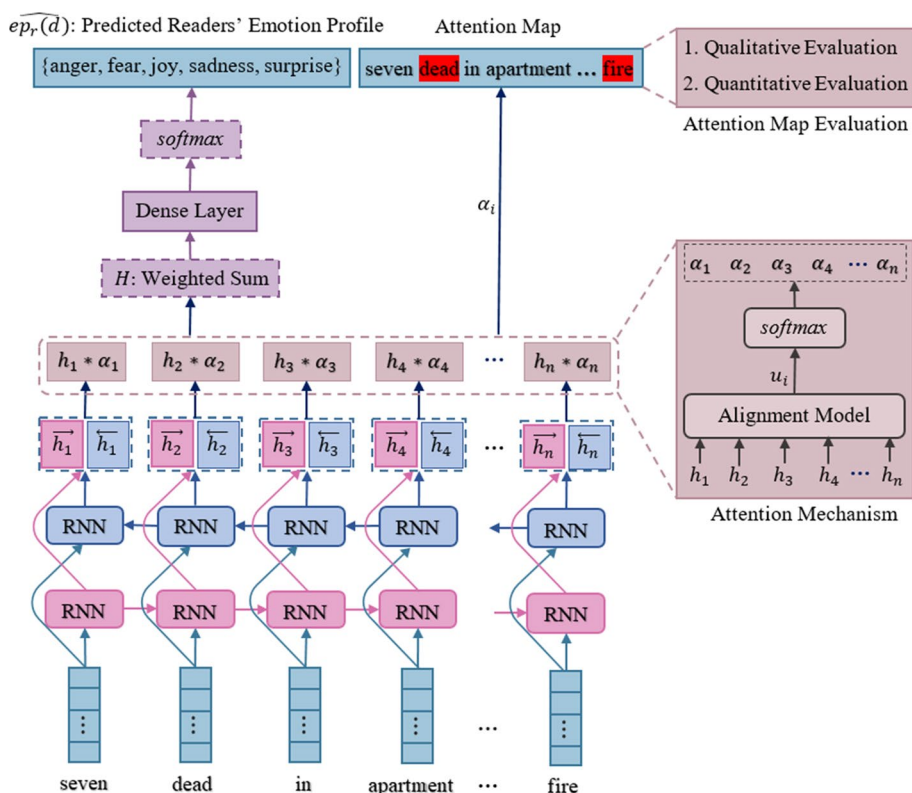


Fig. 1 Detailed sketch of the proposed work

Methodology

To build the readers' emotion detection model, we use one of the prominent RNN based architecture, Bi-LSTM [58], combined with Attention [16]. Our choice of deep learning architecture is oriented towards ensuring model performance as well as ability to investigate model behavior (i.e., interpretability of model). The Bi-LSTM network is capable of learning long-term dependencies without maintaining duplicate context representations [59], and works by performing sequential modeling in both (left to right, and right to left) directions by incorporating past and future context information effectively [17]. The Attention modelling on top of the Bi-LSTM network provides weightage to relevant words in input sequence that highly correlate to our task of prediction. Apart from enhancing overall model performance [19], the use of Attention helps to analyze interpretability of our model towards readers' emotion detection. In particular, it aids our intent towards analyzing how the presence of emotion words and named entities relate to the workings of reader emotion identification. This interpretability analysis objective is attained using explanations precipitated as Attention Maps from the attention layer. A detailed sketch of our technique is illustrated in a self-explanatory manner within Fig. 1.

The Bi-LSTM network is capable of processing sequential inputs from left to right (forward) and from right to left (backward) at the same time to produce contextual information as the output vectors. Let \vec{h}_i be the forward processing hidden layer and \overleftarrow{h}_i be the backward processing hidden layer, concatenated to form a single layer h defined by $[\vec{h}_i; \overleftarrow{h}_i]$. The Bi-LSTM network can be defined as,

$$\vec{h}_l = LSTM(\vec{h}_{l-1}, w_i, \Theta_f) \tag{2}$$

$$\overleftarrow{h}_l = LSTM(\overleftarrow{h}_{l+1}, w_i, \Theta_b) \tag{3}$$

where, Θ_f and Θ_b represent parameters of forward and backward LSTM units, w_i serves as the representation of each word. To learn representations that assign more weightage to those words that contribute significantly to the model’s decision making, we exploit an attention mechanism on top of Bi-LSTM by adopting the popular Attention mechanism proposed by Bahdanau et al. [16]. To implement Attention, initially, we take the last hidden state h_n as a document summary vector Z and process it through an alignment model, which is a feedforward network trained along with the entire model, to produce a scalar value u_i , and later use *softmax* to obtain weights α_i that represents importance of each hidden state h_i .

$$\text{i.e., } u_i = v^\top \tanh(W_h h_i + W_Z Z) \tag{4}$$

$$\alpha_i = \frac{\exp(u_i)}{\sum_{j=1}^n \exp(u_j)} \tag{5}$$

where, $W_h, W_Z \in \mathbb{R}^{a \times b}$ and $v \in \mathbb{R}^a$ are the learnable weight parameters. The final document representation H just before prediction layer is then computed as a weighted sum over h_i and their corresponding weights α_i , denoted as,

$$H = \sum_{i=1}^n \alpha_i h_i \tag{6}$$

This helps to execute the attention mechanism by determining for which words in the source document attention or weightage has to be paid. H is then fed to the output layer, which consists of a single fully connected Multi-Layer Perceptron (MLP) (i.e., dense layer) network capable of producing a normalized distribution of readers’ emotions using a softmax,

$$\widehat{ep_r(d)} = \text{softmax}(\text{MLP}(H)) \tag{7}$$

The loss between $\widehat{ep_r(d)}$ and labelled vector $ep_r(d)$ is propagated back to complete the learning process. Once the model is trained, we empirically evaluate the model on two fronts. First, the accuracy of emotion prediction is evaluated based on how well the predicted emotion distribution reflects the distribution derived from the labels. Second, the attention outputs from documents from a fully-trained network, as indicated, will be qualitatively and quantitatively evaluated to assess model behavior, as outlined in the following section.

Empirical study

We conduct experiments to analyze the performance of Bi-LSTM + Attention model and compare against a number of baselines to illustrate that Bi-LSTM + Attention shows significant performance improvement in detecting reader’s emotions. We then consider

evaluating model behavior with respect to understanding its workings, particularly with a focus on understanding the role of emotion words and named entity mentions. We first describe datasets used in this study, followed by experimental setup and evaluations of model performance and model behavior, with corresponding results and discussion.

Dataset

In our experiments, we utilize three datasets, two *Readers' Emotion News Datasets* (RENh-4k and REN-10k) that we have newly curated, and the SemEval-2007 [25] benchmark dataset.

Readers' emotion news datasets

To procure our two Readers' Emotion News datasets, we use the social news network, Rappler [60] and its award-winning Mood Meter² widget. Mood Meter enables readers to cast their emotion votes towards several categories of emotions (Afraid, Amused, Angry, Annoyed, Don't care, Happy, Inspired, and Sad) and records the total percentage of votes obtained for each emotion. Unlike other sources, we choose Rappler due to its simplicity, popularity, and ease of organizing several news articles under multiple genres and associated emotion profiles. We manually collect only the popular news articles by checking for high emotion votings represented in the Rappler Mood Meter, to ensure that the selected news articles have a high social reach. The detailed information of our two datasets is given below.

RENh-4k: This is a short-text dataset with 4000 news documents and associated readers' emotion profiles. News headlines and associated abstract/snippet are combined to form the documents, and corresponding readers' emotion profiles are obtained from readers' votings on Mood Meter for emotion classes: Afraid, Angry, Happy, Inspired, and Sad. We also assign documents into either of the categories, Health & well-being, Social issues or Others, after manually verifying news genres.

REN-10k: This is an advanced version of RENh-4k, in terms of the number of documents, length of documents, and much diverse set of emotion classes and document genres. This dataset contains 10,272 news documents with corresponding readers' emotion profiles. Here, documents comprise news headlines, abstracts, and news content or full-length news stories without non-textual content like images and videos. Unlike RENh-4k, readers' emotion profiles are collected for a wider set of emotion classes: Afraid, Amused, Angry, Annoyed, Don't care, Happy, Inspired, and Sad. We also assign documents to the categories Business, Entertainment, Lifestyle, Sports, Technology, and Others, by manually verifying genre information available in Rappler. REN-10k documents consist of the whole textual content associated with a particular news article, the average words per document is 533.613, i.e., long-text in nature. Since our study is over short-text documents, we utilize only the news headlines and associated abstracts of REN-10k to form the documents without the associated news content or full-length news stories.

SemEval-2007

SemEval-2007 is a short-text dataset consisting of 1250 documents comprising of news headlines and corresponding emotion scores for the emotion classes Anger, Disgust, Fear, Joy, Sadness, and Surprise, annotated by six readers [25].

² www.web.archive.org/web/20140513012056/http://thenewmedia.com/2012-boomerang-awards-winners/.

Table 1 Dataset statistics after pre-processing

Statistics	REN-10k	RENh-4k	SemEval-2007
Source	Rappler	Rappler	The New York Times, CNN, BBC, Google News
Year span	2014 to 2019	2015 to 2018	–
Length	Short-text (<i>after pre-processing</i>)	Short-text	Short-text
Number of news documents	10,272	4000	1246 (<i>valid documents after pre-processing</i>)
Total number of words	305,160	124,172	6364
Number of unique words	27,749	13,260	3286
Average words per document	29.70	31.043	5.09
Average sentences per document	1.18	1.1875	1.00
Number of annotations	528,327	242,680	6 (<i>annotators</i>)
Mean percentage of votes for each emotion class	Anger: 0.2124 Fear: 0.0658 Joy: 0.4215 Sadness: 0.1399 Surprise: 0.1606	Anger: 0.3388 Fear: 0.1475 Joy: 0.3137 Sadness: 0.0781 Surprise: 0.1218	Anger: 0.1013 Fear: 0.1639 Joy: 0.2860 Sadness: 0.2069 Surprise: 0.2416
Number of articles associated with each emotion class	Anger: 6904 Fear: 4233 Joy: 8917 Sadness: 5972 Surprise: 6431	Anger: 3068 Fear: 1850 Joy: 3267 Sadness: 2489 Surprise: 2312	Anger: 652 Fear: 820 Joy: 786 Sadness: 863 Surprise: 1102

Dataset pre-processing

Given our intent of predicting basic emotions elicited from readers, the first set of pre-processing we perform on datasets is an emotion label mapping from Rappler Mood Meter emotion classes to Paul Ekman's basic emotions [21]. We map *Angry*→*Anger*, *Sad*→*Sadness*, *Afraid*→*Fear*, *Happy*→*Joy* and *Inspired*→*Surprise* and discard other Mood Meter emotion classes such as *Don't care*, *Inspired*, *Amused*, and *Annoyed* by following the methodology proposed by Badaro et al. [30] and Staiano et al. [61]. Since *Disgust* in Ekman's basic emotions do not match with any of the Mood Meter emotion classes, we discard it in our study and maintain rest five basic emotions to preserve common set of labels for all the datasets, as done in [30, 61]. To represent output labels in a better way, as a distribution of five emotions (*anger*, *sadness*, *fear*, *joy*, and *surprise*) we follow a normalization procedure similar to that of Lei et al. [27]. We then perform data cleaning in our datasets by removing noisy or metadata keywords like *report*, *new-review*, *survey*, (*UPDATED*), *Midday-wRa*, etc., that appear several times in the articles. To improve quality of text representation, we also apply generic set of pre-processing techniques including removal of unknown symbols and

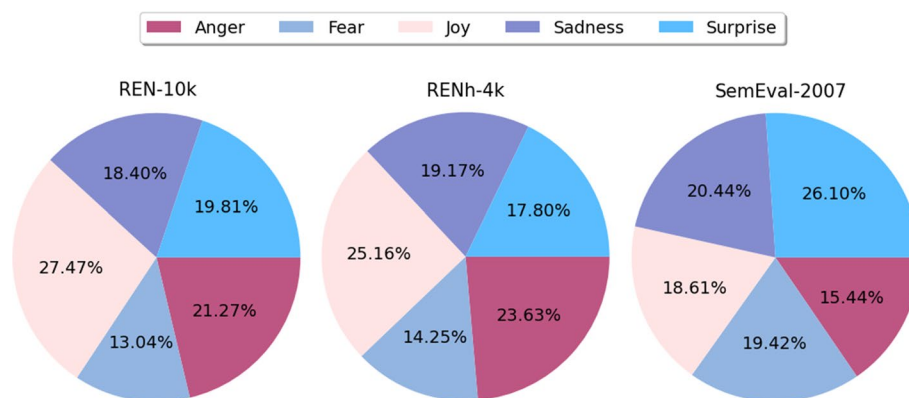


Fig. 2 Distribution of emotions in the datasets

punctuations, and text normalization, using NLTK toolkits³. The detailed statistics of datasets after pre-processing are shown in table 1. Unlike SemEval-2007 which is labeled by six annotators, there is no accurate means to compute number of emotion votings or annotations in Mood Meter, therefore we follow a strategy similar to Guerini et al. [62] to derive the statistics. Figure 2 depicts distribution of emotions in each of the datasets.

Experimental setup and evaluations

We conduct two sets of experiments to evaluate our Bi-LSTM + Attention model for detecting readers' emotions from short-text documents. The first set of experiments focuses on *model performance evaluation* where we compare the performance of our model with several baselines using various coarse-grained and fine-grained evaluation measures. The second set of evaluations focuses on *model behavior analysis* (i.e., interpretability of the model) using the attention maps generated during the predictions. In model behavior analysis, we initially perform an ablation study to identify the impact of attention in predicting readers' emotion profiles, followed by a novel set of qualitative and quantitative evaluation techniques over the attention maps to extensively scrutinize the model's decision making, specifically to realize the role of emotion words and named entities in readers' emotion detection.

Model performance evaluation

To conduct our empirical evaluation, each of the datasets are split into train, validation, and test sets in the ratio 60:20:20 of total dataset volume. To build our Bi-LSTM + Attention model, we embed input documents using different pre-trained word embeddings, Google Word2Vec-300d⁴, Wikipedia2Vec-100d and 200d⁵, and Glove-100d⁶. The *dropout* set to 0.5, Mean Squared Error (*MSE*) as loss function, *Adam* optimizer with

³ <https://www.nltk.org/>.

⁴ <https://code.google.com/archive/p/word2vec/>.

⁵ <https://wikipedia2vec.github.io/wikipedia2vec/pretrained/>.

⁶ <https://nlp.stanford.edu/projects/glove/>.

learning rate 0.0005, batch size 128, $l_2(0.001)$ regularizer, and 100 epochs, are hyper-parameters that can aid reproducibility of our work.

To compare the performance of our Bi-LSTM + Attention model, we implement a set of baselines belonging to the categories deep learning, lexicon based, and classical machine learning (as outlined while discussing related work). Deep learning baselines include the recent state-of-the-art textual emotion detection works and other popular architectures. The lexicon and classical machine learning baselines also include the popular and top-performing state-of-the-art methods. We outline the details of baselines below:

Deep learning baselines

- sent2affect [48]: This is a textual emotion detection method that utilizes transfer learning from an RNN model initially trained for the task of sentiment analysis. Towards reproducing their work faithfully, we use sentiment140⁷ dataset to build the model; the Twitter Sentiment dataset used in their paper was not found in the relevant link provided⁸. We believe sentiment140 is appropriate for usage primarily due to its large size, comprising as much as 1.6 million data objects.
- SS-BED [44]: This is a semantic and sentiment oriented textual emotion detection system, where the same text is subject to two different representations, the semantic representation using word embedding, and the sentiment representation using sentiment specific word embedding proposed in [63].
- Kim's CNN [64]: This work is a popular CNN architecture for text classification. The hyper-parameters used to build this model are given in Appendix.
- Naïve Deep Learning Baselines: Includes the general RNN architectures like GRU [46] and, LSTM and Bi-LSTM used as baselines in certain textual emotion detection works [15, 44, 48]. The hyper-parameters used are given in Appendix.

Lexicon based baselines

- SWAT [11]: SWAT is one of the top ranked systems developed on the shared task, SemEval-2007 Task 14: Affective Text [25]. This supervised system uses predefined sets of emotion words, developed using a unigram model to build emotion annotation of news headlines.
- Emotion Term Model [12]: This is an improved version of the classical Naïve Bayes that incorporates information of emotion rating along with the term independence assumption.
- Synesketech [33]: This is a textual emotion detection system that makes use of a word-level lexicon and an emoticon lexicon, along with a set of heuristic rules.

⁷ <https://www.kaggle.com/kazanova/sentiment140>.

⁸ <https://www.kaggle.com/c/twitter-sentiment-analysis2/data>.

Classical machine learning baselines

- WMD [39]: WMD comprises a textual emotion detection method using Word Mover's Distance feature along with SVM classifier. To reproduce this work faithfully, we use 60% of our corpus for training, 20% for testing, and rest 20% for seed corpus, for the five emotion classes. We use Support Vector Regression (SVR) with multi-output regressor for our multi-target regression problem instead of their SVM classifier.
- Multi-target regression with handcrafted features: We use multiple methods for multi-target regression, with a rich set of features. We describe the features and the models below:
 - TF-IDF Feature [39, 48]: This is a popular and commonly used feature vector indicating Term Frequency (TF) and Inverse Document Frequency (IDF).
 - N-Grams Feature [32, 44]: Towards using the N-Grams feature, we choose N from {1, 2, 3, 4}. For improved efficiency, we utilize Parts-of-Speech tagging to identify and retain only the noun, verb, adverb, and adjectives as they are a prominent source of subjective content [65].
 - General Purpose Emotion Lexicon Features [32]: Total Emotion Count (TEC), Total Emotion Intensity (TEI), Max Emotion Intensity (MEI), Graded Emotion Count (GEC), and Graded Emotion Intensity (GEI), extracted by using a general purpose emotion lexicon, DepecheMood++ [31].
 - Sentiment Word Feature [32, 65]: Combination of two sets of sentiment-oriented features to form a single sentiment word feature. The first set of features capture total number of positive, negative, and neutral words, and the second set computes average positive, negative, and neutral sentiment intensity for a document. We make use of VADER [66], to compute the sentiment features.
 - Embedding Features [44, 63]: Two different types of embeddings, the semantic embeddings which include Word2Vec, GloVe and FastText, and the Sentiment Specific Word Embedding, SSWE_u proposed in [63]. The individual word vectors are averaged to form document vectors for both the embeddings.
 - Multi-target Regression Models: We now describe the multi-target regression models across various families of methods. Based on the problem transformation approach, we implement Multi-output Regressor using Ridge⁹, SVR¹⁰, and GradientBoostingRegressor¹¹. Within the algorithm adaptation approach, we implement a Multi-Layer Perceptron with a single hidden layer of 128 neurons, *ReLU* activation and *l2*(0.001) regularizer, and final output layer with *softmax* activation. Other hyperparameters are *MSE* loss function, *Adam* optimizer with a learning rate 0.0005, batch size set to 64, and 100 *epochs*.

⁹ <https://scikit-learn.org/stable/modules/generated/sklearn.multioutput.MultiOutputRegressor.html#examples-using-sklearnmultioutput-multioutputregressor>.

¹⁰ <https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html>.

¹¹ <https://scikit-learn.org/stable/modules/multiclass.html#multioutput-regression>.

Performance evaluation measures

To measure the effectiveness of readers’ emotion detection, we make use of different coarse-grained and fine-grained evaluation metrics [67]. Coarse-grained measures are useful to understand the correctness of prediction at a binary level, whereas fine-grained measures indicate the nearness of prediction to ground truth. In coarse-grained evaluation, we map regression predictions to a 0/1 classification problem and use Acc@1 (accuracy of top first prediction), a measure that effectively maps to the micro-averaged F1 measure [68]. Acc@1 is popularly used in several textual emotion detection works [12, 13, 27, 69] to measure the performance of a corpus with imbalanced distribution of data. In fine-grained evaluation, we use a set of measures such as AP_{document}, AP_{emotion}, Root Mean Square Error and Wasserstein Distance, which we will describe shortly. AP_{document} and AP_{emotion} are quite popular in textual emotion detection [11, 13, 70] and takes into consideration the correlation between predicted emotion probabilities and ground truth readers’ emotion reactions over the emotions and documents respectively. Our task being formulated as a regression problem uses Root Mean Square Error and Wasserstein Distance that gives a sense of how close (or distant) the predicted emotion probabilities are from the ground truth.

- ◇ Acc@1 [12]: An accuracy measure of the corpus computed by averaging Acc_d@1 of all documents. For a document *d*, Acc_d@1 simply checks whether the top-ranked emotion according to the prediction, (arg max_i ep_r(*d*)[i]), matches the top-ranked emotion according to the label, (arg max_i ep_r(*d*)[i]) i.e.,

$$Acc_d@1 = \begin{cases} 1 & \text{if, } (\arg \max_i ep_r(d)[i] = \arg \max_i \widehat{ep_r(d)}[i]) \\ 0 & \text{else} \end{cases} \tag{8}$$

- ◇ AP_{document} [13]: Average Pearson’s correlation coefficient of the corpus computed by averaging the Pearson’s correlation coefficient P_d of all documents. For each document *d* in the corpus, P_d illustrates correlation between the predicted emotion profile X_d (shorthand for $\widehat{ep_r(d)}$) and ground truth emotion profile Y_d (shorthand for ep_r(*d*)), computed as,

$$P_d = \frac{\sum_{i=1}^{|E|} (X_{d_i} - \bar{X}_d)(Y_{d_i} - \bar{Y}_d)}{(|E| - 1)\sigma_{X_d}\sigma_{Y_d}} \tag{9}$$

where, |E| indicates number of emotion classes, $\bar{X}_d, \bar{Y}_d, \sigma_{X_d}, \sigma_{Y_d}$ are mean and standard deviation of the predicted and ground-truth emotion profiles, respectively. This value may range from [-1, 1], where 1 and -1 indicate perfect positive and perfect negative correlation.

- ◇ AP_{emotion} [13]: Average Pearson’s correlation coefficient of the emotions computed by averaging the Pearson’s correlation coefficient P_e of each emotion category, across all documents. Let A and B be the predicted and ground-truth emotion profiles of an emotion category *e*, then P_e for the emotion category is computed as,

$$P_e = \frac{\sum_{j=1}^{|D|} (A_j - \bar{A})(B_j - \bar{B})}{(|D| - 1)\sigma_A\sigma_B} \tag{10}$$

- ◇ RMSE_D [44]: An error metric for the corpus computed by averaging Root Mean

Square Error $RMSE_d$ of all documents. Let X_d and Y_d be the ground-truth and predicted emotion profiles for document d , then $RMSE_d$ for a document d is computed by,

$$RMSE_d = \sqrt{\frac{\sum_{i=1}^{|E|} (X_d[i] - Y_d[i])^2}{|E|}} \quad (11)$$

- ◇ WD_D [71]: Wasserstein Distance measure is often used to quantify the uncertainty correlated with true error, where true error indicates difference between ground-truth and predicted emotion profiles, X_d and Y_d , respectively. Wasserstein Distance for a document d defined as the infimum for any transport plane is represented as,

$$WD_d(X_d, Y_d) = \inf_{\gamma \sim \pi(X_d, Y_d)} \mathbb{E}_{(x,y) \sim \gamma} [\|x - y\|] \quad (12)$$

where, $\pi(X_d, Y_d)$ is the set of all possible joint probability distribution $\gamma(x, y)$ whose marginals are X_d and Y_d respectively. The average of $WD_d(X_d, Y_d)$ for all documents gives WD_D of the corpus. Lower values indicates good conformance, and thus better performance.

Results and discussion

Table 2 shows the performance of our Bi-LSTM + Attention model and baselines over the REN-10k dataset for various evaluation measures (best result for the entire dataset and best result among baselines in each category are highlighted in boldface for this and later tables). Experimental results show that our model obtains a significant gain¹² of 5.44, 9.04, 6.52, 7.09 and 3.90 percentage points for $Acc@1$, $AP_{document}$, $AP_{emotion}$, $RMSE_D$ and WD_D , respectively, when compared to SS-BED that obtains best results among deep learning baselines and 6.98, 18.43, 21.13, 10.51, and 6.6 percentage points when compared to SWAT and Emotion Term Model that obtains best results among lexicon based baselines. Within the family of problem transformation approaches, we include results of WMD feature with SVR, and linguistic and affective features with Ridge Regression and exclude SVR Regressor and GradientBoostingRegressor since their results are comparatively poor for all the three datasets. Similarly, we tabulate only results of $N = 1$ of N-Grams (unigrams) for which we obtain the best results (a trend similar to [32]). Results illustrate that our model performs well against all other problem transformation baselines and obtains a gain of 9.65, 19.06, 24.49, 8.01, and 4.2 percentage points for $Acc@1$, $AP_{document}$, $AP_{emotion}$, $RMSE_D$, and WD_D , respectively, when compared to best results among problem transformation baselines. Algorithm adaptation baselines show that ANN follows similar trends with improved results than problem transformation approach, where our model even then obtains a gain of 6.75, 13.69, 20.73, 7.21, and 3.97 percentage points for $Acc@1$, $AP_{document}$, $AP_{emotion}$, $RMSE_D$, and WD_D , respectively when compared to best results among algorithm adaptation baselines.

Similar trends are observed for evaluation results over the other two datasets RENh-4k and SemEval-2007. Results of RENh-4k in table 3 demonstrate that for evaluation

¹² by the word *gain* we mean increase in percentage points (↑) for the measures $Acc@1$, $AP_{document}$ and $AP_{emotion}$, and decrease in percentage points (↓) for the measures $RMSE_D$ and WD_D

Table 2 Evaluation results over the REN-10k dataset

Model	Acc@1 (%) \uparrow	AP _{document} \uparrow	AP _{emotion} \uparrow	RMSE _D \downarrow	WD _D \downarrow
Bi-LSTM + Attention (<i>Our Method</i>)	60.55	0.7994	0.5596	0.1500	0.0812
Deep learning baselines					
sent2affect [48]	49.39	0.5716	0.1004	0.2383	0.1298
SS-BED [44]	55.11	0.7090	0.4944	0.2209	0.1202
Kim's CNN [64]	49.03	0.5893	0.1610	0.2332	0.1322
Bi-LSTM [48]	52.80	0.6282	0.4804	0.2215	0.1202
LSTM [9]	52.07	0.6064	0.4581	0.2223	0.1204
GRU	50.17	0.6012	0.2013	0.2329	0.1293
Lexicon based baselines					
SWAT [11]	51.28	0.6151	0.3483	0.2551	0.1472
Emotion Term Model [12]	53.57	0.6023	0.0115	0.3343	0.2520
Synsketch [33]	35.86	0.1632	0.2326	0.2677	0.1664
Problem transformation baselines					
WMD [39]	43.56	0.2366	0.0981	0.3156	0.1480
TF-IDF [39, 48]	49.47	0.6019	0.3133	0.2347	0.1235
N-Grams [32, 44] ($N = 1$)	48.85	0.5331	0.2512	0.2362	0.1251
TEC [32]	50.90	0.6035	0.3133	0.2460	0.1297
TEI [32]	50.90	0.6088	0.3147	0.2301	0.1243
MEI [32]	50.85	0.6029	0.2379	0.2310	0.1255
GEC [32] ($\delta = 0.25$)	50.67	0.6021	0.2765	0.2388	0.1238
GEI [32] ($\delta = 0.25$)	50.63	0.6007	0.2731	0.2392	0.1232
Sentiment word count [32, 65]	50.12	0.6050	0.1939	0.2323	0.1274
SSWE _u [63] ($d = 50$)	49.48	0.5726	0.0714	0.2384	0.1280
GloVe [44] ($d = 100$)	49.63	0.5670	0.0716	0.2390	0.1279
Algorithm adaptation baselines					
TF-IDF [39, 48]	50.17	0.6071	0.2555	0.2303	0.1268
N-Grams [32, 44] ($N = 1$)	50.03	0.5829	0.2173	0.2354	0.1347
TEC [32]	50.51	0.6625	0.3523	0.2257	0.1214
TEI [32]	53.80	0.6516	0.3211	0.2252	0.1209
MEI [32]	49.53	0.5713	0.1859	0.2380	0.1291
GEC [32] ($\delta = 0.25$)	51.24	0.6423	0.2758	0.2285	0.1218
GEI [32] ($\delta = 0.25$)	52.60	0.6163	0.2322	0.2221	0.1269
Sentiment word count [32, 65]	50.36	0.6014	0.1839	0.2331	0.1254
SSWE _u [63] ($d = 50$)	49.44	0.5173	0.0984	0.3751	0.1330
GloVe [44] ($d = 100$)	49.44	0.5169	0.0509	0.3758	0.1334

The best results within each category have been shown in boldface

measures Acc@1, AP_{document}, AP_{emotion}, RMSE_D and WD_D our model achieves a gain of 4.88, 2.02, 4.45, 0.99, and 2.04 percentage points over best results among deep learning baselines, 6.40, 6.41, 10.49, 2.60, and 3.88 percentage points over best results among lexicon based baselines, 6.13, 5.91, 5.22, 1.08, and 0.96 percentage points over best results among problem transformation baselines and 5.75, 4.70, 5.26, 1.05, and 1.23 percentage points over best results among algorithm adaptation baselines, respectively. Table 4 shows results of SemEval-2007 where for the same set of evaluation measures our model achieves a gain of 2.20, 10.01, 4.08, 0.71, and 1.59 percentage points over best results among deep learning baselines, 3.20, 14.98, 15.25, 7.53, and 4.39 percentage points over best results among lexicon based baselines, 7.00, 12.40, 8.71, 3.28, and 2.20 percentage

Table 3 Evaluation results over the RENh-4k dataset

Model	Acc@1 (%) \uparrow	AP _{document} \uparrow	AP _{emotion} \uparrow	RMSE _D \downarrow	WD _D \downarrow
Bi-LSTM + Attention (<i>Our Method</i>)	50.50	0.6499	0.4054	0.2301	0.1220
Deep learning baselines					
sent2affect [48]	36.00	0.4684	0.1047	0.2508	0.1458
SS-BED [44]	45.62	0.5534	0.3609	0.2406	0.1424
Kim's CNN [64]	40.00	0.4775	0.2084	0.2493	0.1585
Bi-LSTM [48]	45.00	0.6297	0.3415	0.2400	0.1465
LSTM [9]	40.13	0.5927	0.3402	0.2559	0.1472
GRU	38.75	0.4860	0.1765	0.2481	0.1443
Lexicon based baselines					
SWAT [11]	43.75	0.5858	0.3005	0.2561	0.1608
Emotion Term Model [12]	44.10	0.5520	0.0102	0.3369	0.2000
Synsketch [33]	31.37	0.1394	0.2423	0.2936	0.1792
Problem transformation baselines					
WMD [39]	35.25	0.3593	0.0289	0.2869	0.1346
TF-IDF [39, 48]	44.37	0.5007	0.3490	0.2440	0.1316
N-Grams [32, 44] ($N = 1$)	42.37	0.5067	0.3009	0.2662	0.1328
TEC [32]	41.12	0.5686	0.3237	0.2410	0.1357
TEI [32]	44.06	0.5908	0.3532	0.2409	0.1316
MEI [32]	40.75	0.5394	0.2574	0.2442	0.1411
GEC [32] ($\delta = 0.25$)	42.75	0.5676	0.3063	0.2410	0.1363
GEI [32] ($\delta = 0.25$)	41.75	0.5602	0.2963	0.2417	0.1365
Sentiment word count [32, 65]	39.25	0.4883	0.1443	0.2492	0.1386
SSWE _u [63] ($d = 50$)	41.50	0.4969	0.1804	0.2483	0.1367
GloVe [44] ($d = 100$)	40.75	0.5108	0.2072	0.2474	0.1327
Algorithm adaptation baselines					
TF-IDF [39, 48]	39.62	0.4630	0.2870	0.2516	0.1489
N-Grams [32, 44] ($N = 1$)	42.75	0.4926	0.2796	0.2456	0.1505
TEC [32]	41.37	0.5701	0.3298	0.2496	0.1356
TEI [32]	42.87	0.6029	0.3528	0.2473	0.1343
MEI [32]	40.12	0.4856	0.2279	0.2488	0.1466
GEC [32] ($\delta = 0.25$)	44.75	0.5726	0.3190	0.2406	0.1359
GEI [32] ($\delta = 0.25$)	41.37	0.5532	0.2934	0.2419	0.1378
Sentiment word count [32, 65]	39.62	0.4846	0.1343	0.2491	0.1425
SSWE _u [63] ($d = 50$)	35.62	0.3080	0.0207	0.4246	0.1376
GloVe [44] ($d = 100$)	35.37	0.2382	0.0920	0.4373	0.1376

The best results within each category have been shown in boldface

points over best results among problem transformation baselines and 3.00, 11.06, 5.13, 3.05, and 2.07 percentage points over best results among algorithm adaptation baselines, respectively.

The overall trends across the results are consistent and suggest that our Bi-LSTM + Attention model performs well on prediction of both highest (Acc@1) and overall (AP_{document} and AP_{emotion}) readers' emotion profiles along with lower values for error and distance metrics over three different datasets. This indicates that the Bi-LSTM + Attention model is able to leverage two-way learning and attention effectively towards identifying readers' emotions. Among several deep learning baselines, SS-BED performs better because we believe it encodes both sentiment and semantic information

Table 4 Evaluation results over the SemEval-2007 dataset

Model	Acc@1 (%) \uparrow	AP _{document} \uparrow	AP _{emotion} \uparrow	RMSE _D \downarrow	WD _D \downarrow
Bi-LSTM + Attention (<i>Our Method</i>)	52.60	0.7140	0.5506	0.1700	0.0915
Deep learning baselines					
sent2affect [48]	37.20	0.3339	0.1075	0.2241	0.1428
SS-BED [44]	50.40	0.6139	0.5098	0.1771	0.1090
Kim's CNN [64]	47.20	0.5437	0.4451	0.1987	0.1200
Bi-LSTM [48]	49.89	0.6007	0.5059	0.1812	0.1074
LSTM [9]	49.20	0.6015	0.5248	0.1842	0.1089
GRU	46.00	0.5673	0.5003	0.2005	0.1098
Lexicon based baselines					
SWAT [11]	46.00	0.4945	0.3981	0.2453	0.1354
Emotion Term Model [12]	49.40	0.5642	0.0167	0.3031	0.1975
Synsketch [33]	35.86	0.3705	0.3570	0.2470	0.1510
Problem transformation baselines					
WMD [39]	40.50	0.1447	0.0459	0.2430	0.1143
TF-IDF [39, 48]	45.60	0.4954	0.4039	0.2080	0.1135
N-Grams [32, 44] ($N = 1$)	45.00	0.4992	0.3931	0.2089	0.1189
TEC [32]	45.20	0.5451	0.4219	0.2028	0.1219
TEI [32]	45.60	0.5900	0.4635	0.2985	0.1228
MEI [32]	45.60	0.4884	0.4071	0.2051	0.1257
GEC [32] ($\delta = 0.25$)	40.80	0.4643	0.3398	0.2113	0.1251
GEI [32] ($\delta = 0.25$)	44.00	0.4416	0.3207	0.2136	0.1291
Sentiment word count [32, 65]	39.04	0.5604	0.3820	0.2089	0.1208
SSWE _u [63] ($d = 50$)	34.56	0.3130	0.1152	0.2300	0.1272
GloVe [44] ($d = 100$)	33.12	0.2605	0.1088	0.2378	0.1152
Algorithm adaptation baselines					
TF-IDF [39, 48]	46.40	0.4799	0.3941	0.2059	0.1206
N-Grams [32, 44] ($N = 1$)	46.80	0.5135	0.4140	0.2027	0.1171
TEC [32]	46.40	0.5639	0.4270	0.2021	0.1204
TEI [32]	49.60	0.6034	0.4993	0.2005	0.1122
MEI [32]	46.40	0.4949	0.4103	0.2062	0.1306
GEC [32] ($\delta = 0.25$)	46.00	0.4861	0.3622	0.2089	0.1229
GEI [32] ($\delta = 0.25$)	46.70	0.4722	0.3531	0.2099	0.1248
Sentiment word count [32, 65]	40.00	0.5732	0.3798	0.2023	0.1193
SSWE _u [63] ($d = 50$)	40.80	0.2071	0.0595	0.4032	0.1641
GloVe [44] ($d = 100$)	42.40	0.2261	0.0777	0.4022	0.1643

The best results within each category have been shown in boldface

to enhance the traditional way of embedding. Transfer learning, in general, gives good results, but on contrary, sent2affect shows low results for sentiment to emotion transfer learning, in our experiments. Our informed guess is that this might be because the source model was built over Twitter data meant specifically for the coarse-grained sentiment classification task, but the target model is meant for an entirely different fine-grained emotion regression task. Whereas in the original implementation of sent2affect, they build both source and target models with similar kinds of Twitter data, both meant for the classification task, which leads to better alignment. In the case of lexicon based baselines, we can observe that SWAT performs well even being an old baseline. We believe that both SWAT and Emotion Term Model could effectively utilize word features

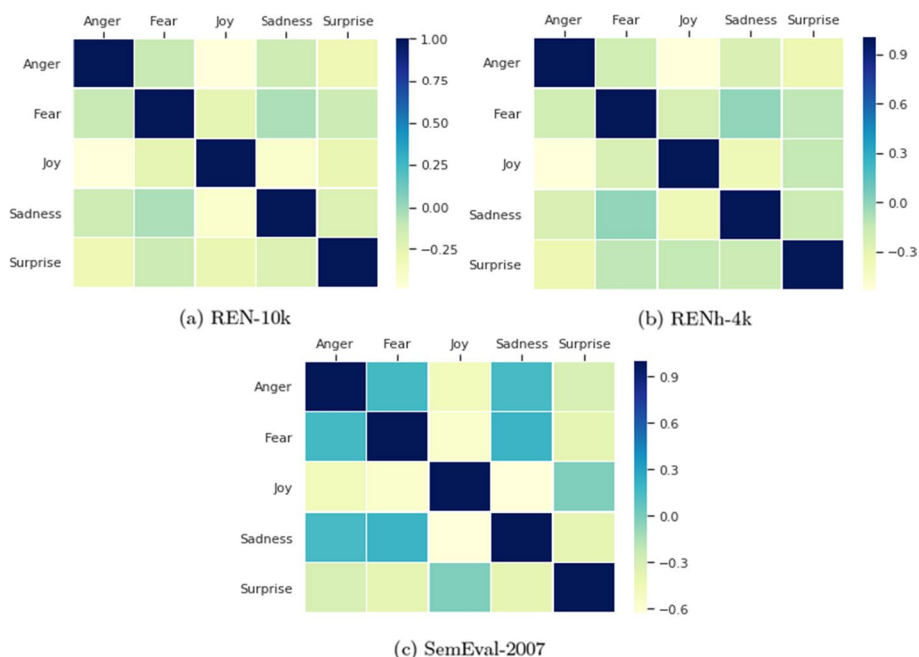


Fig. 3 Emotion profile correlations in the datasets

available within corpora which makes them top performing baselines. On the other hand, Synesketech uses a very generic and non-filtered general-purpose emotion lexicon as the major component (except rule sets), which may be the cause for low results. In machine learning baselines, among various features, affective features, more specifically TEI, outperform traditional linguistic features like TF-IDF and N-Grams in many cases, where TF-IDF, TEC, and MEI are others producing the best results. We also analyze affective features GEC and GEI with three different thresholds of δ , 0.25, 0.5, and 0.75, where we observe degradation in performance with an increase of δ from 0.25 to 0.75, which we believe is due to decreased coverage of emotion words by lexicon, as mentioned in [32].

Performance evaluation across multiple datasets illustrates that SemEval-2007 dataset shows slightly better results than RENh-4k, even with less amount of data. We suppose that SemEval-2007 with labels sourced from up to six annotators is a less complex and better curated dataset. But in the context of our datasets, the minimum number of annotators involved is 242,680 for RENh-4k, and 528,327 for REN-10k, which makes it a complex real-world dataset with several contradictory readers’ votings in ground-truth emotion profiles. To understand the effect of dataset complexity with respect to the number of readers’ annotating a document, we find the degree of correlation between emotions, using Pearson’s correlation coefficient [10], shown in Fig. 3 (dark colors indicate high correlation and light colors indicate low correlation). We can observe several natural correlations in SemEval-2007 such as, *anger* highly correlated to *fear* and *sadness*, but in REN-10k and RENh-4k, a low correlation exists between them. Also, when we observe the correlation between *joy* and *fear* in SemEval-2007, there exists a very low correlation between them, whereas, for REN-10k and RENh-4k, they have comparatively slightly higher correlations. We assume these kinds of irregular and complex patterns

potentially due to noise across a large number of annotators reduce the performance gain in RENh-4k, which is overcome with huge amounts of data in REN-10k producing remarkable gains by allowing to learn the complex patterns.

In addition to the substantial gain observed over various evaluation measures, we statistically evaluate the difference between models by conducting statistical significance tests on paired models in terms of the ideal measures, Acc@1 and RMSE, which are highly capable of representing coarse-grained (i.e., classification) and fine-grained (i.e., regression) characteristics of our task, respectively. We perform McNemar's test over Acc@1 and Kolmogorov-Smirnov test over RMSE to compute the significance between our Bi-LSTM + Attention model and the best baseline using the conventional significance level, i.e., a p-value of 0.05. We obtain statistically significant results corresponding to p-values of 4.79E-11, 3.46E-3, and 8.87E-3 for Acc@1 and 2.96E-19, 1.45E-3, and 3.89E-10 for RMSE, for the three datasets REN-10k, RENh-4k, and SemEval-2007, respectively, which indicates that the results of our Bi-LSTM + Attention model are statistically significant over the best baselines.

Model behavior analysis

Readers' emotions elicited from textual documents may be intuitively expected to be highly oriented towards emotion words and named entities present in the documents. However, such assumptions need to be verified empirically, so they may inform further research into reader emotion detection. In this context, we set our evaluation hypothesis that *key terms that could have helped prediction of readers' emotion profiles in our Bi-LSTM + Attention model are emotion words and named entities present in the documents*. Every prediction of our Bi-LSTM + Attention model produces readers' emotion profiles along with an attention map that highlights key terms (terms which are given weightage by the Attention). Our Bi-LSTM + Attention model enables to analyze the attention maps and hence model behavior (i.e., model's decision making) in the context of readers' emotion detection. Based on our hypothesis, we expect that the attention map of predictions must highlight emotion words and named entities present in the textual document as key terms. Hence, in this section, we devise novel evaluation strategies to computationally represent and validate the hypothesis by initially verifying the necessity of attention mechanism for the task followed by qualitatively and quantitatively analyzing the behavior of our Bi-LSTM + Attention model.

Ablation study over attention layer—uniform attention as the adversary

There is no point in analyzing model behavior to study the impact of emotion words and named entities if attention does not have a reasonable influence on prediction [56]. Hence, to establish the necessity of attention mechanism in our readers' emotion detection task experimented with three different datasets, we adopt a technique similar to ablations studies in machine learning. We study the importance of attention by using *uniform attention as the fall back model* based on observations in [56] that analysis or interpretability of attention stays valid only if it performs as a necessary component in the entire prediction model. For this, we rebuild our model by altering the attention mechanism on top of hidden states, with uniform weights instead of varying weight distributions (a uniform-attention model). This can, in a sense, nullify the effect of attention

Table 5 Comparison with *Uniform Attention as the Adversary* mechanism

Approach	Acc@1 (%)↑	AP _{document} ↑	AP _{emotion} ↑	RMSE _D ↓	WD _D ↓
REN-10k					
Uniform Attention	54.36	0.6963	0.4019	0.2200	0.1125
Bi-LSTM + Attention (<i>Baseline-Our Method</i>)	60.55	0.7994	0.5596	0.1500	0.0812
RENh-4k					
Uniform Attention	46.87	0.6156	0.3515	0.2435	0.1357
Bi-LSTM + Attention (<i>Baseline-Our Method</i>)	50.50	0.6499	0.4054	0.2301	0.1220
SemEval-2007					
Uniform Attention	46.98	0.6490	0.5255	0.2050	0.1105
Bi-LSTM + Attention (<i>Baseline-Our Method</i>)	52.60	0.7140	0.5506	0.1700	0.0915

The best results within each category have been shown in boldface

layer so that we can analyze the model without the influence of attention, and compare it against the model with an attention mechanism, making the study an ablation analysis. Results obtained for *uniform attention as the adversary* experiments for the three datasets are given in Table 5 and is compared against our Bi-LSTM + Attention model, taken as a baseline. The results indicate that our model has noteworthy gains over all the datasets for all evaluation measures. McNemar's test over Acc@1 and Kolmogorov–Smirnov test over RMSE are also computed to analyze the statistical significance between attention enabled (i.e., our Bi-LSTM + Attention model) and uniform-attention (*uniform attention as the adversary*) models. Results illustrate that gains obtained for our Bi-LSTM + Attention model over uniform-attention model are statistically significant with p-values of 34.37E-4, 6.26E-3, and 1.59E-03 for Acc@1 and 1.52E-5, 2.41E-3, and 3.68E-04 for RMSE, for the three datasets REN-10k, RENh-4k, and SemEval-2007, respectively. Thus, the ablation study shows that the attention mechanism in our model significantly influences readers' emotion detection for all three datasets. This provides us confidence that the attention map could contain important information to verify our hypothesis with respect to emotion words and named entities.

Qualitative evaluation

Qualitative evaluation is conducted by manually investigating the presence of key terms in attention maps based on the hypothesis that what the model specifically looks for giving a weightage in the task of readers' emotion detection, are emotion words and named entities. Table 6 shows two sets of attention maps generated through our model with their associated ground truth (ep_r) and predicted (\hat{ep}_r) emotion profiles. Color intensities over the words in attention maps indicate weightage associated with the words, i.e., dark red indicates high weightage for the words, whereas light red indicates less weightage. In the first set of attention maps, we include samples whose predicted emotion profiles are very near to ground truth, hence we categorize them as correct predictions. The first attention map among the correct predictions set shows that a high-intensity weightage is given to the word 'attack' and then to the words 'hiding' and 'threats' with a slight weightage decay, which explains the nearness of predicted emotion profiles to ground truth. That is, higher values are seen to peak around the emotions, *fear*, *sadness* and *anger*, for both predicted and ground truth emotion profiles, which undoubtedly showcases the intimate relationship between attention recognized words and

Table 6 Sample attention maps (ep_r : ground truth, \hat{ep}_r : predicted)

Document Attention Map	Emotion profiles for [anger, fear, joy, sadness, surprise]
Correct predictions	
Teacher in hiding after attack on Islam Stirs Threats	$ep_r = [0.149, 0.436, 0.000, 0.413, 0.000]$ $\hat{ep}_r = [0.233, 0.430, 0.026, 0.311, 0.000]$
Bad weather slows S.Korean search Russian ship	$ep_r = [0.120, 0.040, 0.000, 0.840, 0.000]$ $\hat{ep}_r = [0.102, 0.012, 0.001, 0.790, 0.091]$
Women protest Pakistan demolition	$ep_r = [0.339, 0.122, 0.000, 0.245, 0.292]$ $\hat{ep}_r = [0.330, 0.210, 0.003, 0.280, 0.170]$
126 students suffer food poisoning Makati. Most of the students who experienced and dizziness stomach pains after ingesting snacks bought from school canteen, have been discharged from the hospital.	$ep_r = [0.173, 0.062, 0.062, 0.617, 0.086]$ $\hat{ep}_r = [0.188, 0.086, 0.071, 0.517, 0.136]$
Ines Fernandez mother others. Nanay Ines shining example of women who are empowering mothers in rural areas to take better care of their health and wellbeing through proper nutrition and education .	$ep_r = [0.000, 0.000, 0.271, 0.000, 0.729]$ $\hat{ep}_r = [0.007, 0.040, 0.353, 0.057, 0.550]$
Warriors destroyed by Blazers as Lillard scores 51. Golden State falls to 48-5, with all 5 defeats coming on the road	$ep_r = [0.326, 0.000, 0.413, 0.174, 0.087]$ $\hat{ep}_r = [0.318, 0.001, 0.465, 0.182, 0.032]$
Incorrect predictions	
Greek police hunt embassy attackers	$ep_r = [0.551, 0.252, 0.046, 0.149, 0.000]$ $\hat{ep}_r = [0.187, 0.277, 0.080, 0.301, 0.152]$
Personal health : for teenagers , car is the danger zone	$ep_r = [0.000, 0.494, 0.000, 0.221, 0.284]$ $\hat{ep}_r = [0.109, 0.229, 0.104, 0.349, 0.207]$
The sweet tune of an anniversary	$ep_r = [0.000, 0.000, 1.000, 0.000, 0.000]$ $\hat{ep}_r = [0.016, 0.026, 0.545, 0.247, 0.167]$
33 killed in Central Luzon since start of election gun ban . Most of them were killed during shootout with authorities while evading checkpoints , says Central Luzon police director Chief Superintendent Joel Napoleon Coronel.	$ep_r = [0.290, 0.570, 0.000, 0.000, 0.140]$ $\hat{ep}_r = [0.378, 0.250, 0.076, 0.244, 0.050]$
PH Air Force to welcome 2 fighter jets . The squadron of 12 brand new fighter jet will be completed within the year , according to Air Force spokesman Colonel Antonio Francisco.	$ep_r = [0.000, 0.011, 0.915, 0.000, 0.074]$ $\hat{ep}_r = [0.106, 0.068, 0.586, 0.088, 0.149]$
Liberia's last Ebola patient discharged. Almost 24,000 people have been infected with the virus since 2013 December	$ep_r = [0.000, 0.000, 0.500, 0.000, 0.500]$ $\hat{ep}_r = [0.130, 0.183, 0.189, 0.382, 0.108]$

emotions. Similarly, many other attention maps in the correct predictions set show a substantial weightage for emotion words; for example, the words ‘pain’, ‘suffer’, ‘poisoning’ in the fourth attention map and the association of predicted emotion profiles with emotion *sadness*. Also, the fifth attention map highlights words such as ‘shining’, ‘better’, ‘care’, ‘empowering’, which may be the reason to predict high intensities for emotions *surprise* and *joy*. Next, we observe weightage associated with named entities in the attention maps. In the correct predictions set we identify that many named entities like ‘Korean’, ‘Pakistan’, ‘Lillard’, ‘Ines Fernandez’, etc., are highlighted with varying weightages. For example, in the sixth attention map, we believe that the word ‘Lillard’ (name of an American basketball player) may also have influenced to produce high-intensity for emotion *joy* in some readers and *anger* in others, besides other words with an attention weightage. From the perspective of such qualitative analyses, we infer that attention gives high weightage to emotion words and nearly so to the named entities for the task of readers’ emotion detection.

In contrast to the first set of correct predictions, we include a few random samples from incorrect predictions and their attention maps also, in Table 6 as the second set. By incorrect predictions, we mean to refer to predictions that are far away from the patterns

Table 7 Emotion Lexicon coverage (in percentages)

Dataset	DepecheMood++	EmoWordNet	NRC-Affect Intensity Lexicon
REN-10k	80.26	53.03	9.66
RENh-4k	88.11	67.13	13.67
SemEval-2007	94.69	86.50	20.28

of ground truth emotion profiles. Here too, we can observe that attention maps highlight a few emotion terms and named entities such as ‘danger’, ‘killed’, ‘Antonio’, etc., but it has missed most of the relevant ones. For example, in the first attention map among incorrect predictions, attention gives zero weightage to the word ‘attackers’, which we believe has enough power to predict high intensities for the emotions *anger*, *fear*, and *sadness*, similar to ground truth emotion profile. Apart from these kinds of exclusion of key terms (i.e., emotion words and named entities), we also identify that most of the incorrect predictions assign high weightage to many words like ‘says’, ‘year’, ‘almost’, ‘since’, etc. Hence, we believe that a major reason for the increase in gap between predicted and ground truth emotion profiles is due to the exclusion of emotion words and named entities, and instead assigning high weightage to many less significant words in the document. This correlation between focus on emotion words and named entities, with measures of performance further reasserts the value of emotion words and named entities in the readers’ emotion detection task.

Quantitative evaluation

From above mentioned qualitative evaluations, we observe that attention maps give weightage mostly to emotion words and named entities. Hence in this section, we bring forth a novel set of evaluation measures to quantify the presence of emotion words and named entities in predictions. Therefore, apart from machine attention maps generated internally by our model, we devise external attention maps that can highlight emotion words and named entities by leveraging external information (e.g., lexicons). To generate external lexicon-based attention maps, we initially identify three popular emotion lexicons, NRC-Affect Intensity Lexicon [29], EmoWordNet [30] and DepecheMood++ [31], and compute lexicon coverage for unique words in the datasets used in our study, results are shown in table 7. We can observe that both DepecheMood++ and EmoWordNet gives better coverage, hence we choose these two lexicons for our quantitative studies the details of which will follow soon. Further, to identify named entities, we use an external tool, specifically the Named Entity Recognizer (NER) from spaCy¹³. The construction of the extrinsic attention maps will be evident through their definitions that follow.

Definition 1 (DAM) This is the internal Document Attention Map produced by the model for each input document (from the attention layer), represented as a vector with intensity values or weightage associated with each word, which indicates the attention received by that word during prediction. If the weightage of words in the attention map

¹³ <https://spacy.io/>.

is continuous then it is called a continuous attention map; DAM is generally a continuous representation. But if the weightage is either 0 or 1, indicating the presence or absence of attention for a certain word, then it may be called a binary attention map.

Definition 2 (EmoNE-EAM) This External Attention Map is independent of the DAM (and thus, the BiLSTM – Attention method) and is generated with the help of an emotion lexicon (we use [30, 31]) and Named Entity Recognizer (we use NER from spaCy). To create EmoNE-EAM, we read each word in the document sequentially and set attention weightage of the words to a boolean value 1 if it is an element of emotion lexicon or NER, else set to 0. This map will be a binary representation that indicates only the presence of emotion words and named entities in the document.

Definition 3 (EmoNE-HAM) This Hybrid Attention Map is generated by considering only the words that have non-zero attention weightage in DAM, and thus blends the information from across the model-generated map and the external information from lexicons. This map can have continuous or binary representations. We create continuous EmoNE-HAM similar to EmoNE-EAM, but boolean values in EmoNE-EAM are replaced by the weightages in DAM, provided the words have non-zero weightages in DAM. In the case of binary EmoNE-HAM, instead of adding DAM weightages to the attention map, we set the value as 1. That is, EmoNE-HAM represents only emotion words and named entities in a document that is recognized by DAM.

The above attention maps provide us a convenient platform to measure the impact of emotion words and named entities in the prediction. For computational convenience, we accomplish this by contrasting the extent of deviation between the EAM (external attention map) and the HAM (hybrid attention map). We quantitatively measure the impact of emotion words and named entities in prediction by finding the overlap between the HAM and EAM, using three measures, namely, behavioral similarity, word similarity, and word probability.

- ◇ Behavioral Similarity: Motivated by [22], we compute the behavioral similarity of corpus D as the average pair-wise similarity between EmoNE-HAM and EmoNE-EAM for all the documents, given as,

$$\text{BehSim}_D = \frac{1}{D} \sum_{d=1}^{|D|} \text{AUC}(\text{EmoNE-HAM}_d, \text{EmoNE-EAM}_d) \quad (13)$$

where, EmoNE-HAM_d is a continuous attention map vector and EmoNE-EAM_d is a binary attention map, for each document d . AUC ranges between 0 and 1, with perfect similarity given by 1, no similarity by 0.5, and negative similarity by 0 [22]. A high behavioral similarity will occur in cases where the model gives high intensity weightage for emotion words and named entities.

- ◇ Word Similarity: This measures the similarity between attention maps in the

Table 8 Quantitative evaluation results

Dataset	DepecheMood++	EmoWordNet
Behavioural similarity scores		
REN-10k	0.8829	0.8497
RENh-4k	0.7096	0.6988
SemEval-2007	0.8092	0.8040
Word similarity scores		
REN-10k	0.8296	0.8010
RENh-4k	0.6851	0.6606
SemEval-2007	0.8203	0.7919
Word probability scores		
REN-10k	0.9043	0.8901
RENh-4k	0.7648	0.7205
SemEval-2007	0.8981	0.8624

context of cosine angle projected in a multi-dimensional space. Word similarity score WordSim_D for corpus D is computed by averaging cosine similarities¹⁴ of binary EmoNE-HAM and EmoNE-EAM for all the documents.

$$\text{WordSim}_D = \frac{1}{|D| - |D'|} \sum_{d=1}^{|D|-|D'|} \cos(\text{EmoNE-HAM}_d, \text{EmoNE-EAM}_d) \quad (14)$$

where, $|D'|$ indicates the number of documents that don't have any emotion words or named entities. This measure computes the similarity between emotion words and named entities identified by our attention mechanism on one side, and total emotion words and named entities present in the document on the other.

- ◇ **Word Probability:** A measure that uses boolean intersection between binary EmoNE-HAM and EmoNE-EAM to quantify how much emotion words and named entities are identified by the attention mechanism during prediction, among the total number of emotion words and named entities present in the document. Unlike the previous similarity scores, this measure is represented in probabilities. We compute word probability WordProb_D for corpus D by averaging word probabilities of all the documents.

$$\text{WordProb}_D = \frac{1}{|D| - |D'|} \sum_{d=1}^{|D|-|D'|} \frac{\sum(\text{EmoNE-EAM}_d \cap \text{EmoNE-HAM}_d)}{\sum(\text{EmoNE-EAM}_d) + \lambda} \quad (15)$$

where, $\lambda = 1$ only if $\text{EmoNE-EAM} = 0$, and $\lambda = 0$ if $\text{EmoNE-EAM} \neq 0$.

Experimental results of quantitative evaluation of model behavior for the three datasets are illustrated in Table 8. In the case of behavioral similarity, the highest score of 0.8829 is observed for the model trained on REN-10k dataset, and the lowest score of 0.6988 is observed for the model trained on RENh-4k (this is still greater than 0.5), indicating a good amount of similarity between the model generated and external attention maps. Word similarity scores also show a good amount of similarity between these attention

¹⁴ <https://deeppai.org/machine-learning-glossary-and-terms/cosine-similarity>.

maps, where the model trained on REN-10k obtains the highest score of 0.8296 and the model trained on RENh-4k obtains the lowest score of 0.6606. For word probability, the highest score of 0.9043 is observed for the model trained on REN-10k and the lowest score of 0.7205 is observed for the model trained on RENh-4k, which indicates that attention captures a significant amount of emotion words and named entities to make the predictions. Promising and consistent results observed for the datasets over all the evaluation measures for both lexicons indicate that our attention mechanism highly relies on emotion words and named entities for predicting readers' emotion profiles. The general trend of scores decaying from REN-10k to RENh-4k reflects the prediction performances of the model trained on these datasets as shown in Tables 2, 3, 4, i.e., REN-10k gives best prediction results whereas, RENh-4k gives comparatively low prediction results in model performance analysis, and hence their quantitative behavior evaluation scores.

Conclusion

In this paper, we explored a Bi-LSTM + Attention model to predict the emotion profiles of readers' towards short-text documents. The simple design of our method ensures generalizable operation and allows a detailed evaluation of model behavior to draw reusable insights, especially that oriented towards assessing the interpretable nature of attention mechanism for the task of readers' emotion detection. To perform the experiments we procured two new readers' emotion news datasets, REN-10k and RENh-4k that can aid extensive studies in the future. Apart from our datasets, we also utilize the benchmark SemEval-2007 dataset. Our first phase of experiments for *model performance evaluations* using various coarse-grained and fine-grained measures shows that Bi-LSTM + Attention outperforms the baselines belonging to different categories of emotion detection including deep learning, lexicon based, and classical machine learning, with remarkable gains. We also performed *model behavior evaluations* using a novel set of qualitative and quantitative methods to interpret the workings of the attention mechanism; these studies firmly establish that emotion words and named entities significantly influence readers' emotion detection.

Future directions

Given that our study establishes emotion words significantly influence readers' emotion detection, we are considering to explore the scope of emotion-specific embedding with the combinations of Bi-LSTM + Attention and transformer based language models. Further, we are considering to develop an improved version of our dataset (REN-20k) to handle dataset complexities due to contradictory emotions provided for the documents depending on readers'/annotators votings. There is also a large scope for further evaluation with a completely human-generated attention map (as in [22]), apart from the model generated and external attention maps, to build better computational models.

Appendix

Hyper-parameters used to build the deep learning baselines, Kim's CNN [64], GRU, LSTM [9], and Bi-LSTM [48] are provided in Table 9 to aid reproducibility.

Table 9 Hyper-parameters of the deep learning baselines

Parameters	Kim's CNN	GRU	LSTM	Bi-LSTM
Filter size	3, 4 and 5	–	–	–
Number of filters	100	–	–	–
Number of RNN Stack	–	1	1	1
Neurons in Stack	–	100	100	100
Embedding	Pre-trained GloVe	Pre-trained GloVe	Pre-trained GloVe	Pre-trained GloVe
Embedding dimension	100	100	100	100
Regularizer	$l2(0.01)$	$l2(0.01)$	$l2(0.001)$	$l2(0.001)$
Dropout	0.5	0.25	0.5	0.5
Loss	MSE	MSE	MSE	MSE
Optimiser	<i>Adam</i>	<i>Adam</i>	<i>Adam</i>	<i>Adam</i>
Learning rate	0.0005	0.005	0.0005	0.0005
Dense layer activation	<i>softmax</i>	<i>softmax</i>	<i>softmax</i>	<i>softmax</i>
Batch size	64	64	128	128
Epoch	100	100	100	100

Abbreviations

NLP	Natural language processing
CNN	Convolutional neural network
RNN	Recurrent neural network
LSTM	Long Short-term memory
Bi-LSTM	Bidirectional long short-term memory
REN	Readers' emotion news
REnh	Readers' emotion news headlines
SVM	Support vector machine
WMD	Word mover's distance
GRU	Gated recurrent unit
MLP	Multi-layer perceptron
MSE	Mean squared error
Adam	Adaptive moment estimation
SVR	Support vector regression
ANN	Artificial neural network
TF	Term frequency
IDF	Inverse document frequency
POS	Parts-of-speech
TEC	Total emotion count
TEI	Total emotion intensity
MEI	Max emotion intensity
GEC	Graded emotion count
GEI	Graded emotion intensity
VADER	Valence aware dictionary and sEntiment reasoner
SSWE	Sentiment specific word embedding
RMSE	Root mean square error
WD	Wasserstein distance
DAM	Document attention map
NER	Named entity recognizer

Acknowledgements

The first author (AK) wishes to dedicate this work to the ever-loving memory of his father Ayyappan K. The authors thankfully acknowledge the project interns Renjitha Rajendran and Shonima Sanil, Department of Information Technology Kannur University, Athira Biju and Sruthi S Kumar, Department of Computer Science Mahatma Gandhi University, and Amrutha Praseeth, Diya Rajan and Rahul Das H, the postgraduate students of Department of Computer Science University of Calicut, who have been involved in dataset procurement. The fifth author (MPG) would like to thank the Department of Science and Technology (DST) of the Government of India for financial support through the Women Scientist Scheme-A (WOS-A) for Research in Basic/Applied Science under the Grant SR/WOS-A/PM-62/2018. The authors thankfully acknowledge the popular leading digital media company RAPPLER for allowing to procure news data along with associated emotions from their online portal that very relevantly helped to conduct this research.

Author contributions

AK, DP, and LVL initiated the work. AK and DP played key roles in conceptualization. AK, DP, and SSA designed the algorithm and experimental workflow. AK and MPG obtained the datasets for the research, implemented and managed the

coding. The rich experience of DP was instrumental in refining the work. The manuscript was collaboratively authored by AK and MPG under the supervision of DP and LVL. All authors contributed to the editing and proofreading. All authors read and approved the final manuscript.

Funding

This research was not supported by any funded Project.

Availability of data and materials

The datasets procured during the current study are available from the authors on reasonable request and also publicly available at <https://dcs.uoc.ac.in/cida/resources/ren-10k.html> for REN-10k and at <https://dcs.uoc.ac.in/cida/resources/renh-4k.html> for RENh-4k.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

The authors give the Publisher the permission to publish the work.

Competing interests

The authors declare that they have no competing interests.

Received: 30 September 2021 Accepted: 11 April 2022

Published online: 20 June 2022

References

- Lin KH-Y, Chen H-H. Ranking reader emotions using pairwise loss minimization and emotional distribution regression. In: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Honolulu, Hawaii; 2008. p. 136–144. <https://aclanthology.org/D08-1015>.
- Lin KH-Y, Yang C, Chen H-H. Emotion classification of online news articles from the reader's perspective. In: 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, vol 1; 2008. p. 220–226. <https://doi.org/10.1109/WIIAT.2008.197>.
- Chang Y-C, Chu C-H, Chen CC, Hsu W-L. Linguistic template extraction for recognizing reader-emotion. In: International Journal of Computational Linguistics & Chinese Language Processing, Volume 21, Number 1, June 2016; 2016. <https://aclanthology.org/O16-2002>.
- Bhowmick PK, Basu A, Mitra P. Reader perspective emotion analysis in text through ensemble based multi-label classification framework. *Comput Inf Sci*. 2009;2(4):64–74. <https://doi.org/10.5539/cis.v2n4p64>.
- Ye L, Xu R-F, Xu J. Emotion prediction of news articles from reader's perspective based on multi-label classification. In: 2012 International Conference on Machine Learning and Cybernetics, vol 5; 2012. p. 2019–2024. <https://doi.org/10.1109/ICMLC.2012.6359686>. IEEE.
- Xu R, Ye L, Xu J. Reader's emotion prediction based on weighted latent dirichlet allocation and multi-label k-nearest neighbor model. *J Comput Inf Syst*. 2013;9(6):2209–16.
- Cabrera-Diego LA, Bessis N, Korkontzelos I. Classifying emotions in stack overflow and jira using a multi-label approach. *Knowl-Based Syst*. 2020;195:105633. <https://doi.org/10.1016/j.knosys.2020.105633>.
- Rao Y, Li Q, Mao X, Wenyin L. Sentiment topic models for social emotion mining. *Inf Sci*. 2014;266:90–100. <https://doi.org/10.1016/j.ins.2013.12.059>.
- Krebs F, Lubascher B, Moers T, Schaap P, Spanakis G. Social Emotion Mining Techniques for Facebook Posts Reaction Prediction. In: Proceedings of the 10th International Conference on Agents and Artificial Intelligence (ICAART), vol 1. SciTePress, INSTICC; 2018. p. 211–220. <https://doi.org/10.5220/0006656002110220>.
- Tang D, Zhang Z, He Y, Lin C, Zhou D. Hidden topic-emotion transition model for multi-level social emotion detection. *Knowl-Based Syst*. 2019;164:426–35. <https://doi.org/10.1016/j.knosys.2018.11.014>.
- Katz P, Singleton M, Wicentowski R. SWAT-MP: the SemEval-2007 systems for task 5 and task 14. In: Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007). Association for Computational Linguistics, Prague, Czech Republic; 2007. p. 308–313. <https://aclanthology.org/S07-1067>.
- Bao S, Xu S, Zhang L, Yan R, Su Z, Han D, Yu Y. Mining social emotions from affective text. *IEEE Trans Knowl Data Eng*. 2011;24(9):1658–70. <https://doi.org/10.1109/TKDE.2011.188>.
- Liang W, Xie H, Rao Y, Lau RY, Wang FL. Universal affective model for readers' emotion classification over short texts. *Expert Syst Appl*. 2018;114:322–33. <https://doi.org/10.1016/j.eswa.2018.07.027>.
- Dong R, Peng O, Li X, Guan X. Cnn-svm with embedded recurrent structure for social emotion prediction. In: 2018 Chinese Automation Congress (CAC); 2018. p. 3024–3029. <https://doi.org/10.1109/CAC.2018.8623318>. IEEE.
- Liu Z-X, Zhang D-G, Luo G-Z, Lian M, Liu B. A new method of emotional analysis based on CNN-BiLSTM hybrid neural network. *Clust Comput*. 2020;23:2901–13. <https://doi.org/10.1007/s10586-020-03055-9>.
- Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate; 2014. arXiv preprint [arXiv:1409.0473](https://arxiv.org/abs/1409.0473).
- Jang B, Kim M, Harerimana G, Kang S-U, Kim JW. Bi-Lstm model to increase accuracy in text classification: Combining word2vec cnn and attention mechanism. *Appl Sci*. 2020. <https://doi.org/10.3390/app10175841>.
- Mishra RK, Urolagin S, Jothi JAA, Neogi AS, Nawaz N. Deep learning-based sentiment analysis and topic modeling on tourism during covid-19 pandemic. *Front Comput Sci*. 2021. <https://doi.org/10.3389/fcomp.2021.775368>.

19. Kardakis S, Perikos I, Grivokostopoulou F, Hatzilygeroudis I. Examining attention mechanisms in deep learning models for sentiment analysis. *Appl Sci*. 2021. <https://doi.org/10.3390/app11093883>.
20. Guan X, Peng Q, Li X, Zhu Z. Social emotion prediction with attention-based hierarchical neural network. In: 2019 IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), vol 1; 2019. p. 1001–1005. <https://doi.org/10.1109/IAEAC47372.2019.8998031>. IEEE.
21. Ekman P. Basic emotions. In: *Handbook of Cognition and Emotion*, Chap. 3, John Wiley & Sons, Ltd.; 1999. p. 45–60. <https://doi.org/10.1002/0470013494.ch3>.
22. Sen C, Hartvigsen T, Yin B, Kong X, Rundensteiner E. Human attention maps for text classification: do humans and neural networks focus on the same words? In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online; 2020. p. 4596–4608. <https://doi.org/10.18653/v1/2020.acl-main.419>.
23. Mohammad SM, Bravo-Marquez F. WASSA-2017 shared task on emotion intensity. In: *Proceedings of the Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA)*, Copenhagen, Denmark; 2017. p. 34–49. <https://doi.org/10.18653/v1/W17-5205>.
24. Mohammad S, Bravo-Marquez F, Salameh M, Kiritchenko S. Semeval-2018 task 1: Affect in tweets. In: *Proceedings of the 12th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, New Orleans, Louisiana; 2018. p. 1–17. <https://doi.org/10.18653/v1/S18-1001>.
25. Strapparava C, Mihalcea R. SemEval-2007 task 14: Affective text. In: *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*. Association for Computational Linguistics, Prague, Czech Republic; 2007. p. 70–74. <https://aclanthology.org/S07-1013>.
26. Li X, Rao Y, Xie H, Liu X, Wong T-L, Wang FL. Social emotion classification based on noise-aware training. *Data Knowl Eng*. 2019;123:101605. <https://doi.org/10.1016/j.datak.2017.07.008>.
27. Lei J, Rao Y, Li Q, Quan X, Wenyan L. Towards building a social emotion detection system for online news. *Futur Gener Comput Syst*. 2014;37:438–48. <https://doi.org/10.1016/j.future.2013.09.024>.
28. Bostan LAM, Kim E, Klinger R. GoodNewsEveryone: A corpus of news headlines annotated with emotions, semantic roles, and reader perception. In: *Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France; 2020. p. 1554–1566. <https://aclanthology.org/2020.lrec-1.194>.
29. Mohammad S. Word affect intensities. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA), Miyazaki, Japan; 2018. <https://aclanthology.org/L18-1027>.
30. Badaro G, Jundi H, Hajj H, El-Hajj W. EmoWordNet: Automatic expansion of emotion lexicon using English WordNet. In: *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*. Association for Computational Linguistics, New Orleans, Louisiana; 2018. p. 86–93. <https://doi.org/10.18653/v1/S18-2009>.
31. Araque O, Gatti L, Staiano J, Guerini M. Depechemood++: a bilingual emotion lexicon built through simple yet powerful techniques. *IEEE Trans Affect Comput*. 2019. <https://doi.org/10.1109/TAFFC.2019.2934444>.
32. Bandhakavi A, Wiratunga N, Padmanabhan D, Massie S. Lexicon based feature extraction for emotion text classification. *Pattern Recogn Lett*. 2017;93:133–42. <https://doi.org/10.1016/j.patrec.2016.12.009>.
33. Krcadinac U, Pasquier P, Jovanovic J, Devedzic V. Synesketch: an open source library for sentence-based emotion recognition. *IEEE Trans Affect Comput*. 2013;4(3):312–25. <https://doi.org/10.1109/T-AFFC.2013.18>.
34. Mulki H, Bechikh Ali C, Haddad H, Babaoğlu I. Tw-StAR at SemEval-2018 task 1: Preprocessing impact on multi-label emotion classification. In: *Proceedings of The 12th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, New Orleans, Louisiana; 2018. p. 167–171. <https://doi.org/10.18653/v1/S18-1024>.
35. Muljono, Winarsih NAS, Supriyanto C. Evaluation of classification methods for indonesian text emotion detection. In: *2016 International Seminar on Application for Technology of Information and Communication (Isemantic)*; 2016. p. 130–133. <https://doi.org/10.1109/ISEMANTIC.2016.7873824>. IEEE.
36. S AD, S R, Rajendram SM, T T M. SSN MLRG1 at SemEval-2018 task 1: Emotion and sentiment intensity detection using rule based feature selection. In: *Proceedings of The 12th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, New Orleans, Louisiana; 2018. p. 324–328. <https://doi.org/10.18653/v1/S18-1048>.
37. Urologin S. Sentiment analysis, visualization and classification of summarized news articles: a novel approach. *Int J Adv Comput Sci Appl*. 2018. <https://doi.org/10.14569/IJACSA.2018.090878>.
38. Urologin S, Thomas S. 3d visualization of sentiment measures and sentiment classification using combined classifier for customer product reviews. *Int J Adv Comput Sci Appl*. 2018. <https://doi.org/10.14569/IJACSA.2018.090508>.
39. Ren F, Liu N. Emotion computing using word mover’s distance features based on ren_cecps. *PLoS ONE*. 2018;13(4):1–17. <https://doi.org/10.1371/journal.pone.0194136>.
40. Xu H, Lan M, Wu Y. ECNU at SemEval-2018 task 1: Emotion intensity prediction using effective features and machine learning models. In: *Proceedings of The 12th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, New Orleans, Louisiana; 2018. p. 231–235. <https://doi.org/10.18653/v1/S18-1035>.
41. Tsoumakas G, Katakis I. Multi-label classification: an overview. *Int J Data Warehousing Mining (IJDWIM)*. 2007;3(3):1–13.
42. Wang Y, Feng S, Wang D, Yu G, Zhang Y. Multi-label Chinese microblog emotion classification via convolutional neural network. In: *Asia-Pacific Web Conference*, Springer; 2016. p. 567–580.
43. Ge S, Qi T, Wu C, Huang Y. THU_NGN at SemEval-2019 task 3: Dialog emotion classification using attentional LSTM-CNN. In: *Proceedings of the 13th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, Minneapolis, Minnesota, USA; 2019. p. 340–344. <https://doi.org/10.18653/v1/S19-2059>.
44. Chatterjee A, Gupta U, Chinnakotla MK, Srikanth R, Galley M, Agrawal P. Understanding emotions in text using deep learning and big data. *Comput Hum Behav*. 2019;93:309–17. <https://doi.org/10.1016/j.chb.2018.12.029>.
45. Shrivastava K, Kumar S, Jain DK. An effective approach for emotion detection in multimedia text data using sequence based convolutional neural network. *Multimedia Tools Appl*. 2019;78(20):29607–39. <https://doi.org/10.1007/s11042-019-07813-9>.

46. Du P, Nie J-Y. Mutux at SemEval-2018 task 1: Exploring impacts of context information on emotion detection. In: Proceedings of The 12th International Workshop on Semantic Evaluation. Association for Computational Linguistics, New Orleans, Louisiana; 2018. p. 345–349. <https://doi.org/10.18653/v1/S18-1052>.
47. Li M, Dong Z, Fan Z, Meng K, Cao J, Ding G, Liu Y, Shan J, Li B. ISCLAB at SemEval-2018 task 1: UIR-miner for affect in tweets. In: Proceedings of The 12th International Workshop on Semantic Evaluation. Association for Computational Linguistics, New Orleans, Louisiana; 2018. p. 286–290. <https://doi.org/10.18653/v1/S18-1042>.
48. Kratzwald B, Ilić S, Kraus M, Feuerriegel S, Prendinger H. Deep learning for affective computing: text-based emotion recognition in decision support. *Decis Support Syst.* 2018;115:24–35. <https://doi.org/10.1016/j.dss.2018.09.002>.
49. Wang C, Wang B, Xiang W, Xu M. Encoding syntactic dependency and topical information for social emotion classification. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR'19. Association for Computing Machinery, New York, NY, USA; 2019. p. 881–884. <https://doi.org/10.1145/3331184.3331287>.
50. Srivastava RA, Deepak G. Piren: prediction of intermediary readers' emotion from news-articles. In: *Data Science and Security*, Springer, Singapore; 2021. p. 122–130.
51. Mou X, Peng Q, Sun Z, Wang Y, Li X, Bashir MF. A deep learning framework for news readers' emotion prediction based on features from news article and pseudo comments. *IEEE Trans Cybern.* 2021. <https://doi.org/10.1109/TCYB.2021.3112578>.
52. Rathnayaka P, Abeysinghe S, Samarajeewa C, Manchanayake I, Walpola MJ, Nawaratne R, Bandaragoda T, Alahakoon D. Gated recurrent neural network approach for multilabel emotion detection in microblogs; 2019. arXiv preprint [arXiv:1907.07653](https://arxiv.org/abs/1907.07653).
53. Yu J, Marujo L, Jiang J, Karuturi P, Brendel W. Improving multi-label emotion classification via sentiment classification with dual attention transfer network. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Brussels, Belgium; 2018. p. 1097–1102. <https://doi.org/10.18653/v1/D18-1137>.
54. Baziotis C, Nikolaos A, Chronopoulou A, Kolovou A, Paraskevopoulos G, Ellinas N, Narayanan S, Potamianos A. NTUA-SLP at SemEval-2018 task 1: Predicting affective content in tweets with deep attentive RNNs and transfer learning. In: Proceedings of The 12th International Workshop on Semantic Evaluation. Association for Computational Linguistics, New Orleans, Louisiana; 2018. p. 245–255. <https://doi.org/10.18653/v1/S18-1037>.
55. Lertvittayakumjorn P, Toni F. Human-grounded evaluations of explanation models for text classification. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Association for Computational Linguistics, Hong Kong, China; 2019. p. 5195–5205. <https://doi.org/10.18653/v1/D19-1523>.
56. Wiegrefe S, Pinter Y. Attention is not not explanation. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Association for Computational Linguistics, Hong Kong, China; 2019. p. 11–20. <https://doi.org/10.18653/v1/D19-1002>.
57. Vashishth S, Upadhyay S, Tomar GS, Faruqi M. Attention interpretability across nlp tasks; 2019. arXiv preprint [arXiv:1909.11218](https://arxiv.org/abs/1909.11218).
58. Schuster M, Paliwal KK. Bidirectional recurrent neural networks. *IEEE Trans Signal Process.* 1997;45(11):2673–81. <https://doi.org/10.1109/78.650093>.
59. Liang D, Zhang Y. Ac-blstm: asymmetric convolutional bidirectional lstm networks for text classification; 2016. arXiv preprint [arXiv:1611.01884](https://arxiv.org/abs/1611.01884).
60. Rappler: Philippine & World News: Investigative Journalism: Data: Civic Engagement: Public Interest. Accessed 20 February 2022. <https://www.rappler.com/>.
61. Staiano J, Guerini M. Depeche mood: a lexicon for emotion analysis from crowd annotated news. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Association for Computational Linguistics, Baltimore, Maryland; 2014. p. 427–433. <https://doi.org/10.3115/v1/P14-2070>.
62. Guerini M, Staiano J. Deep feelings: A massive cross-lingual study on the relation between emotions and virality. In: Proceedings of the 24th International Conference on World Wide Web. WWW '15 Companion. Association for Computing Machinery, New York, NY, USA; 2015. p. 299–305. <https://doi.org/10.1145/2740908.2743058>.
63. Tang D, Wei F, Yang N, Zhou M, Liu T, Qin B. Learning sentiment-specific word embedding for Twitter sentiment classification. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Baltimore, Maryland; 2014. p. 1555–1565. <https://doi.org/10.3115/v1/P14-1146>.
64. Kim Y. Convolutional neural networks for sentence classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, Doha, Qatar; 2014. p. 1746–1751. <https://doi.org/10.3115/v1/D14-1181>.
65. Suharshala R, Anoop K, Lajish VL. Cross-domain sentiment analysis on social media interactions using senti-lexicon based hybrid features. In: 2018 3rd International Conference on Inventive Computation Technologies (ICICT). IEEE, Coimbatore, India; 2018. p. 772–777. <https://doi.org/10.1109/ICICT43934.2018.9034272>.
66. Hutto C, Gilbert E. Vader: a parsimonious rule-based model for sentiment analysis of social media text. *Proc Int AAAI Conf Web Social Media.* 2014;8(1):216–25.
67. Strapparava C, Mihalcea R. Learning to identify emotions in text. In: Proceedings of the 2008 ACM Symposium on Applied Computing. SAC '08. Association for Computing Machinery, New York, NY, USA; 2008. p. 1556–1560. <https://doi.org/10.1145/1363686.1364052>.
68. Manning CD, Raghavan P, Schütze H. Introduction to information retrieval. Cambridge University Press; 2008. <https://books.google.co.in/books?id=t1PoSh4uwVcC>.
69. Rao Y, Li Q, Wenyan L, Wu Q, Quan X. Affective topic model for social emotion detection. *Neural Netw.* 2014;58:29–37. <https://doi.org/10.1016/j.neunet.2014.05.007>.
70. Rao Y, Xie H, Li J, Jin F, Wang FL, Li Q. Social emotion classification of short text via topic-level maximum entropy model. *Inf Manag.* 2016;53(8):978–86. <https://doi.org/10.1016/j.im.2016.04.005>.

71. Ghoshal B, Tucker A. Estimating uncertainty and interpretability in deep learning for coronavirus (covid-19) detection; 2020. arXiv preprint [arXiv:2003.10769](https://arxiv.org/abs/2003.10769).

Publisher's Note

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

Submit your manuscript to a SpringerOpen[®] journal and benefit from:

- ▶ Convenient online submission
- ▶ Rigorous peer review
- ▶ Open access: articles freely available online
- ▶ High visibility within the field
- ▶ Retaining the copyright to your article

Submit your next manuscript at ▶ [springeropen.com](https://www.springeropen.com)
