ESTeR: Combining Word Co-occurrences and Word Associations for Unsupervised Emotion Detection

Sujatha Das Gollapalli*, **Polina Rozenshtein***, **See-Kiong Ng**
Institute of Data Science, National University of Singapore, Singapore
{idssdg, idspoli, seekiong}@nus.edu.sg

Abstract

Accurate detection of emotions in user-generated text was shown to have several applications for e-commerce, public well-being, and disaster management. Currently, the state-of-the-art performance for emotion detection in text is obtained using complex, deep learning models trained on domain-specific, labeled data. In this paper, we propose **E**motion-**S**ensitive **Te**xt**R**ank (*ESTeR*), an unsupervised model for identifying emotions using a novel similarity function based on random walks on graphs. Our model combines large-scale word co-occurrence information with word-associations from lexicons avoiding not only the dependence on labeled datasets, but also an explicit mapping of words to latent spaces used in emotion-enriched word embeddings. Our similarity function can also be computed efficiently. We study a diverse range of datasets including recent tweets related to COVID-19 to illustrate the superior performance of our model and report insights on public emotions during the on-going pandemic.

1 Introduction

Human beings are known to perceive and feel various, highly-nuanced emotions, expressed both in spoken and written texts. Modeling emotions in user-generated content has been shown to benefit domains such as commerce, public health, and disaster management (Bollen et al., 2011b; Neppalli et al., 2017; Hu et al., 2018; Pamungkas, 2019). E.g., emotion cues from social media posts were used to identify depression and PTSD (Deshpande and Rao, 2017; Aragón et al., 2019) and for personalizing chatbots to improve user satisfaction (Wei et al., 2019).

Recent studies list as many as 154 human emotions (Smith, 2015). However, most researchers in Psychology have largely agreed on a set of basic emotions such as *anger*, *fear*, *disgust*, *sadness*, *surprise*, and *happiness* (Ekman, 2016) and showed that complex emotions can be expressed using this basic set (Ekman, 1992; Plutchik, 2001). For example, Plutchik uses combinations, intensity, and opposites of basic emotions for capturing the higher-order emotions. That is, *annoyance* and *rage* can be viewed as the less or more intense forms of *anger*, and *anticipation* is the opposite of *surprise*. Thus, most recent studies on automatic emotion detection use Ekman's or Plutchik's sets of 6 or 8 emotions, respectively (Mohammad et al., 2018; Liu et al., 2019).

Existing models for automatic emotion identification in user-generated texts typically use supervised learning techniques. The state-of-the-art emotion detection performance on tweets, news articles, blogs, reviews, and TV-show transcripts is obtained using complex, deep learning architectures that combine a range of features including terms, embeddings, and domain-specific aspects such as emojis, as well as human-generated lexicons of emotion-word associations (Chatterjee et al., 2019; Zahiri and Choi, 2018; Mundra et al., 2017; Abdul-Mageed and Ungar, 2017; Köper et al., 2017). Much manual effort is involved in collecting annotated data for a given domain and fine-tuning domain-specific models.

Other auxiliary works enabling emotion detection can be placed under two complementary directions. The first one is lexicon development for emotions via manual vocabulary labeling or automatic generation, for example, based on similarity to a set of seed words (Mohammad and Turney, 2013; Araque et al., 2019). The second direction uses a latent space of embeddings to compare sentences with emotion lexicons (Xu et al., 2015; Savigny and Purwarianti, 2017). Compiling a lexicon of a high quality and coverage is a labor-intensive task,

*Equal contribution from both authors.

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 1043–1056
November 16 - 20, 2020.

and even when automation and crowdsourcing is involved, a close manual control is required. As for latent space representations, the embedding model must include sufficient information about the underlying emotions, obtained, e.g., from the lexicons or labeled datasets (Agrawal et al., 2018; Xu et al., 2018; Tang et al., 2014).

Both embeddings and lexicons enable basic techniques for unsupervised emotion prediction– for example, by using word embeddings similarities (Kim et al., 2010) or overlap between lexicon words and input text (Araque et al., 2019). Considering the abundance of user-generated texts on the current-day Web with its ever-changing topics (for example, "COVID-19 lockdown"), we argue that it is desirable to develop *advanced unsupervised models that detect emotions accurately across domains, offer a probabilistic explanation for the predicted emotions, while not depending on large quantities of labeled data*. These desirables comprise our precise objectives in this paper.

We present **E**motion-**S**ensitive **Te**xt**R**ank (*ESTeR*) and its variants as our similarity functions that use word graphs for scoring input texts with reference to a given set of emotions. *ESTeR* is designed based on the following two observations: (1) For a given language, words expressing emotions are fairly stable across domains (Agrawal et al., 2018). For example, the same words ("This is absurd...") may be used to express *anger* (emotion) regarding a product on an e-commerce website as well as in a tweet related to a goverment policy. (2) Word-occurrence graphs are known to capture contextual and latent language information and were successfully used in various NLP tasks (Mihalcea and Tarau, 2004; Yan et al., 2013; Chen and Kao, 2015; Kong et al., 2016).

We make the following contributions:

- For identifying emotions in textual content, we propose similarity functions that incorporate word co-occurrence information from large-scale, publicly-available text corpora and word associations from lexicons. Our novel similarity functions are based on random walks on word graphs and score an input text with respect to a given emotion.
- Next, we formally show the relation between the proposed similarity functions and Personalized PageRank (Haveliwala et al., 2003). In addition, we provide a computational method based on solving a linear system of equations to compute our similarity functions efficiently at the dataset level, rather than per instance.
- We present experiments illustrating the superior performance of our models on five recent, publicly-available datasets for emotion detection (Klinger et al., 2018; Liu et al., 2019).
- Finally, we showcase our proposed model on a newly-collected dataset of COVID-19 tweets by highlighting various interesting aspects of public emotions during the current pandemic.

In the next section (Section 2), we present our scoring framework for emotion detection along with derivations on how to compute our solution efficiently. In Section 3, we summarize datasets and experiments illustrating the performance of our proposed model. In Section 4, we demonstrate anecdotally, the effectiveness of model on COVID-19 tweets. Finally, we present closely-related work in Section 5 and conclusions in Section 6.

2 Methods

2.1 Preliminaries

Given an input text (alternately referred to as a "sentence" in this paper for ease), d, and a set of emotions, $\mathcal{E}$, the objective of the emotion detection task is to identify a subset of emotions from $\mathcal{E}$ to be assigned as labels for d. This objective translates into constructing a score function $s : \mathcal{D} \times \mathcal{E} \rightarrow \mathbb{R}_{\geq 0}$, where $\mathcal{D}$ is a dataset of n sentences.

Similar to previous unsupervised models (Kim et al., 2010), we would like to leverage the information from the emotion lexicons: a set of words $\mathcal{L}(e)$ which have known binary or continuous association with the emotions $e \in \mathcal{E}$. Vocabulary $\mathcal{V}$ of size m is the union of all of words (in lexicon, dataset, and the corpus used to generate our graph-based model, to be explained shortly).

Let x_d and x_e be, respectively, the vector representations of a sentence d and a lexicon of emotion e. We use binary bag-of-words column vectors of length $|\mathcal{V}|$. The matrix $D \in \mathbb{R}^{m \times n}$ represents the dataset with each column corresponding to a sentence vector x_d for some $d \in \mathcal{D}$, whereas each column of the emotions matrix $E \in \mathbb{R}^{m \times |\mathcal{E}|}$ is a vector representation x_e of some emotion $e \in \mathcal{E}$.

The score function $s(d, e)$ is typically defined as a similarity function between the vector representations of a sentence d and emotion e (Seyeditabari et al., 2018). For example, the commonly-used cosine similarity function is given by:

$$s_{cos}(d, e) = cos(x_d, x_e) = \frac{x_d^T x_e}{\|x_d\|_2 \, \|x_e\|_2}. \quad (1)$$

To mitigate problems due to sparsity of lexicons that may result in insufficent overlap between a sentence and the emotion vectors, previous works have employed latent spaces for representing sentences and lexicons. These spaces can be obtained through matrix factorization approaches (Kim et al., 2010) or more recently through neural embeddings (Polignano et al., 2019). For these models the corresponding scoring function $s_{lat}(d, e)$ can be written as

$$cos(Mx_d, Mx_e) = \frac{x_d^T M^T M x_e}{\|Mx_d\|_2 \|Mx_e\|_2}, \quad (2)$$

$M \in \mathbb{R}^{h \times m}$ is the embeddings matrix, $h \ll m$.

2.2 *ESTeR*: Our Proposed Scoring Function

Latent-space based similarity functions show relatively improved performance (Polignano et al., 2019). However, previous works have highlighted the shortcomings of using general latent space representations for specific tasks and often labeled data is used to fine-tune latent representations within supervised models (Seyeditabari and Zadrozny, 2017; Yeh et al., 2017). Therefore, we would like to avoid an explicit mapping into a latent space by turning to the classical notion of random walk-based graph similarity and look for functions of the shape:

$$s(d, e) = \frac{x_d^T P x_e}{norm(x_d, x_e)}, \quad (3)$$

where $norm$ is some appropriate normalization for x_d and/or x_e .

In deriving random-walk based similarity functions on word graphs, we need a transition matrix whose entries represent probabilities of moving from one word to another. Similar to previous works (Mihalcea and Tarau, 2004), we use a co-occurrence matrix $A \in \mathbb{R}_{\geq 0}^{m \times m}$ derived from some general corpus, e.g., Wikipedia, where $A(i, j)$ is the number of times words i and j appear in the same text window. The entries of A are row-normalized to convert A to a stochastic matrix.

In the random walk with restarts model, we assume that at each step of the random walk, the walker proceeds with moving to another word according to the transition matrix with probability $\alpha \in [0, 1]$ and stops with the probability $1 - \alpha$ (Haveliwala, 2003; Yazdani and Popescu-Belis, 2010; Duan et al., 2018). The resulting matrix P can be expressed as:

$$P = (1 - \alpha) \sum_{k=0}^{\infty} \alpha^k A^k,$$

where k is the walk length.

Since our goal is to measure similarity of a sentence to a lexicon, we need to allow the walk to restart only inside the sentence vocabulary. That is, a random walker restarts at any, chosen at random, word $w \in \mathcal{V}$ in d. With a uniform distribution, each word can be chosen with a probability $1/\|x_d\|_1$. This translates into $\|x_d\|_1$ normalization in Equation 3. On the other hand, we would like to reach *any* word in the lexicon, thus the probabilities to reach each particular word in the lexicon are aggregated as a sum without any normalization. Therefore, $norm(x_d, x_e) = \|x_d\|_1$ and the final formula for $s(d, e)$ (we denote it as *ESTeR*(d, e), **E**motion **S**ensitive **T**ext**R**ank) is:

$$ESTeR(d, e) = (1 - \alpha) \frac{x_d^T}{\|x_d\|_1} \sum_{k=0}^{\infty} \alpha^k A^k x_e. \quad (4)$$

ESTeR(d, e) has a clear probabilistic interpretation as the probability that a random walk with restarts in a sentence d ends in the lexicon e.

Note that, the probability that a random walker stops at a word $w \in \mathcal{V}$ restarting from the words of a sentence d is given by:

$$PPR(x_d, w) = (1 - \alpha) \frac{x_d^T}{\|x_d\|_1} \sum_{k=0}^{\infty} \alpha^k A^k e_w,$$

where e_w is a one-hot vector with 1 at the position corresponding to the word w.

Such $PPR(x_d, w)$ is a classic Personalized (Haveliwala et al., 2003) or Topic-Sensitive (Haveliwala, 2003) PageRank score of a word w for a personalization (topic) vector $\frac{x_d^T}{\|x_d\|_1}$. That is,

$$ESTeR(d, e) = PPR(x_d)^T x_e, \quad (5)$$

where $PPR(x_d) \in \mathbb{R}^{m \times 1}$ is a Personalized PageRank vector.

2.3 Computation

Our objective is to compute *ESTeR* for a dataset efficiently. Using matrix representations for a dataset (D) and emotions (E), the Formula 4 can be written as:

$$ESTeR(\mathcal{D}, \mathcal{E}) = (1 - \alpha) D_n^T \sum_{k=0}^{\infty} \alpha^k A^k E,$$

where D_n is a l_1 column-normalized matrix D. *ESTeR*$(\mathcal{D}, \mathcal{E})$ is a matrix of size $n \times |\mathcal{E}|$, where each element *ESTeR*(d, e) is the score of a document d in emotion e. Using Neumann series (see e.g., (Naylor and Sell, 2000)), it can be further written

as:

$$ESTeR(\mathcal{D}, \mathcal{E}) = (1-\alpha)D_n^T(I-\alpha A)^{-1}E.$$

If we calculate *ESTeR*(d, e) naïvely using available methods, we would have to run the PageRank algorithm for each sentence. To avoid this we first solve a linear system:

$$(I-\alpha A)Z = (1-\alpha)E$$

for $Z \in \mathbb{R}^{m\times|\mathcal{E}|}$. We can then calculate the final dot-product with D as: *ESTeR*$(\mathcal{D}, \mathcal{E}) = D_n^T Z$. The total time complexity of this method is $O(|\mathcal{E}|LA(m) + mult(\mathcal{D}_n, Z))$. $LA(m)$ is the cost of solving a linear system of size m and generally takes $O(m^3)$, but for the cases of sparse matrices, such as ours, can run in quadratic to almost linear time in practice (Zlatev, 1991). $mult(\mathcal{D}_n, Z)$ is the time complexity of matrix multiplication, which is $O(m|\mathcal{E}|n)$ in general, but for the multiplication of a sparse $(\mathcal{D}_n)$ and dense (Z) matrices, the complexity can be reduced to $O(nnz(\mathcal{D}_n)|\mathcal{E}|)$, where $nnz(\mathcal{D}_n)$ is the number of non-zero entries in matrix $\mathcal{D}_n$ (Zlatev, 1991). $nnz(\mathcal{D}_n)$ can be estimated as $a \cdot n$ where a is average sentence length. Discarding a and $|\mathcal{E}|$ as constants, the computational complexity is $O(LA(m) + n)$.

Note that, once the system is solved and matrix Z is obtained, we can estimate scores of any new sentences on the fly in linear time of the sentence length, similar to supervised predictive models.

If computed naïvely, even using the popular power method (Arasu et al., 2002) for PageRank computation requires $O(m^2)$ per iteration, thus computing Equation 5 for the whole dataset $\mathcal{D}$ takes $O(nm^2I + n \cdot nnz(E)m)$, I is the iterations number. Estimating the lexicon size as m/b for $b > 1$ and discarding constants results in complexity of $O(nm^2)$.

2.4 Variants

We consider a couple of variations of our *ESTeR* scoring function to enable other probabilistic interpretations of scoring texts with respect to emotions.

ESTeR:LexNorm. The first variant incorporates the normalization on lexicon vectors as:

$$\textit{ESTeR:LexNorm}(\mathcal{D}, \mathcal{E}) = (1-\alpha)D_n^T(I-\alpha A)^{-1}E_n,$$

Here, E_n is column-normalized E by ℓ_1 norm. Since lexicons (particularly auto-generated lexicons) can be large and some emotions have richer word-associations, normalization has the effect of balancing the sizes of the lexicons and contributions of each word. This variant, therefore, captures the probability that a random walk with restarts starts in the sentence and ends in the lexicon, if the starting and ending words $u \in \mathcal{V}(d)$ and $v \in \mathcal{L}$ are chosen uniformly at random.

ESTeR:Lex2Sent. The second variant reverses the intuition for *ESTeR* by capturing the probability that the random walk starts in the lexicon and ends in the sentence and is given by:

$$\textit{ESTeR:Lex2Sent}(\mathcal{D}, \mathcal{E}) = (1-\alpha)E_n^T(I-\alpha A)^{-1}D_n.$$

This variant therefore score sentences based on how well they reflect the lexicon.

2.5 Baselines for Comparisons

Since techniques for unsupervised emotion detection are lacking, we formulate our baselines based on the two resource types created for this task.

For the first set of baselines, we directly use the recent emotion-enriched word embeddings from *ewe-uni300* (Agrawal et al., 2018), *emo2vec100*[1] (Xu et al., 2018), and *sswe-u50*[2] (Tang et al., 2014) to represent sentence and emotion vectors. The similarity is computed using Equation 2. Unlike general word embeddings (Pennington et al., 2014), emotion-enriched embeddings use supervision of some form to capture the "emotion similarity/dissimilarity" between words in a latent space.

The second set of baselines incorporates coverage in emotion lexicons by using Equation 1 to compute the similarity between the sentence and emotion vectors. We use *EmoLex* (Mohammad and Turney, 2013) and *DepecheMood* (Staiano and Guerini, 2014), two recent lexicons that are also in many supervised emotion detection models (Mohammad et al., 2018; Liu et al., 2019). In the next section, we refer to the baseline techniques using the resource names.

3 Experiments and Results

3.1 Datasets

Several annotated datasets are publicly-available for studying emotion detection (Klinger et al., 2018). For our empirical evaluation, we choose the most recent datasets that are labeled using 6 (Ekman's set) or 8 (Plutchik's set) prime emotions. Plutchik's set which is the larger of the two sub-

[1] https://github.com/pxuab/emo2vec_wassa_paper

[2] http://ir.hit.edu.cn/~dytang/paper/sswe/embedding-results.zip

Dataset	anger	anticipation	disgust	fear	joy	sadness	surprise	trust	size
SemEval2018	3960	1527	4020	1848	4319	3233	566	553	10516
SSEC	1390	739	440	274	815	414	177	520	3320
DENS	1304	1019	74	1412	1264	1401	362	1156	7991
TEC	1527	-	760	2505	8140	3829	3803	-	20564
CrowdFlower	110	-	2325	8445	13030	5157	2182	-	31233

Table 1: Dataset size and categorical breakdown.

Lexicon	anger	anticipation	disgust	fear	joy	sadness	surprise	trust	size
EmoLex	1247	839	1058	1476	689	1191	534	1231	14181
DepecheMood	114983	-	110298	92837	168478	115204	129997	-	187940

Table 2: Lexicon size and categorical breakdown.

sumes the Ekman's set and comprises of the emotions: *joy, trust, fear, surprise, sadness, anticipation, anger, and disgust*. Our experimental datasets are briefly summarized below and in Table 1.

SemEval2018 is a dataset of tweets from 2016 and 2017 collected using affective query terms and manually-annotated using crowdsourcing. This dataset was used previously for the "SemEval-2018 Affect in Tweets" challenge task (Mohammad et al., 2018).

SSEC or Stance Sentiment Emotion Corpus is a dataset of stance and sentiment tweets from 2016 annotated for emotions by Schuff et al (2017). Using hashtag keywords, the collected tweets represent users' stances towards a given target topic such as *Climate Change*, *Feminist Movement* and other topics (Mohammad et al., 2017).

DENS (Liu et al., 2019) is a recent dataset containing passages of classic and modern narratives with lengths between 40 and 200 tokens. During labeling, the label *trust* is substituted by *love* so that the labelers could recognize *trust* better in romantic context. We substitute *love* back with *trust* to match our emotion labels.

TEC (Mohammad, 2012) is a dataset of general tweets collected in 2012 by using Ekman's set of emotions as hashtags (e.g., #anger) with the objective to test if the hashtag corresponds to the label.

CrowdFlower[3] is a dataset of general tweets, provided by Microsoft's Cortana Intelligence Gallery. Since this dataset was labeled with 13 non-standard emotional categories, we used the mapping proposed by Klinger el al. (2018) to obtain labels for our 8 emotions.

[3] https://data.world/crowdflower/sentiment-analysis-in-text

Table 1 summarizes the main characteristics of the datasets. Note that all datasets are gathered from Twitter platform, except for *DENS*. Overall, our choice of datasets comprises the most recent datasets available for the emotion detection task (Klinger et al., 2018). Tweet datasets are representative of the abundant, diverse, and ever-changing content on Twitter whereas the recently-collected *DENS* has narrative texts from literature and fan-fiction websites. Together they comprise a diverse collection of datasets to evaluate our proposed unsupervised methods. All datasets except *TEC* permit multi-labeling. The median number of labels for all datasets is 1 except for *SemEval2018* where it is 2.

3.2 Resources and Measures

ESTeR computation depends on two resources: the lexicons providing emotion-word association information and the graph containing word co-occurrence information. The lexicons *EmoLex* and *DepecheMood* described in Section 2.5 are used for *ESTeR* variants as well.

For co-occurrence matrices, we experimented with the following corpora: *Wiki* is based on the Wikipedia dump of text articles collected in Feb 2020 comprising of 39K words and 1.7M non-zero entries, *Twitter* is based on the dataset of tweets (Go et al., 2009) contains 17.7K tokens and 1.3M non-zero entries, and *Combined* is the co-occurrence matrix for the combined corpora with 47.7K tokens and 1.9M non-zero entries.[4]

For computing *ESTeR*, standard BLAS[5] implementations for Linear Algebra subproblems was

[4] Further details on these resources and datasets are included in the Appendix A due to space limitations. The Python 3 implementations of the methods, and experimentation scripts are available at https://github.com/nusids/ester.

[5] http://www.netlib.org/blas/

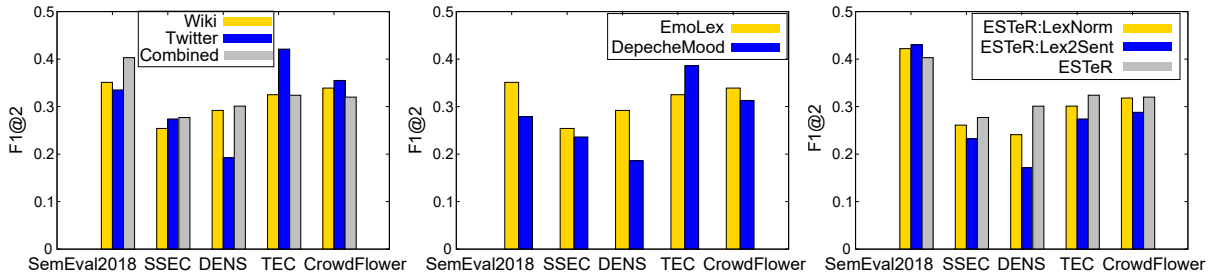

Figure 1: (left) Effect of co-occurrence matrix on *ESTeR*; (middle) Effect of different lexicons; (right) *ESTeR* is compared against *ESTeR:LexNorm* and *ESTeR:Lex2Sent*.

used. For an indicative runtime, we can calculate *ESTeR* scores for *SemEval2018* dataset with *Combined* matrix and *EmoLex* lexicon in 11 minutes in total. [6]

Following previous works on multi-label emotion detection, we present our results using Jaccard Accuracy and $F1$ measure evaluated for top-k predictions, referred to as Jaccard@k and $F1@k$, with k set to $1, 2$. That is, if $L(d)$ indicates the set of correct labels for document d, and $L'(d)$ the predicted set, the measures are given as:

$$Jaccard = \frac{1}{|\mathcal{D}|} \sum_{d \in \mathcal{D}} \frac{|L(d) \cap L'(d)|}{|L(d) \cup L'(d)|},$$

$$F1 = \frac{1}{|\mathcal{D}|} \sum_{d \in \mathcal{D}} \frac{2 \cdot P(d) \cdot R(d)}{P(d) + R(d)}.$$

where $P(d)$ and $R(d)$ refer to the precision and recall for d respectively (Manning et al., 2008):

$$P(d) = \frac{|L(d) \cap L'(d)|}{|L'(d)|}, \; R(d) = \frac{|L(d) \cap L'(d)|}{|L(d)|}.$$

3.3 Experimental results

Effect of co-occurrence matrices on *ESTeR*: In the leftmost plot of Figure 1, we show the effect of using the different co-occurrence matrices from *Wiki*, *Twitter* and *Combined* on our similarity function. Since our contention is that word co-occurrence graphs incorporate the latent information required for emotion detection, the richer and more representative the corpus is, the better *ESTeR* performs for the emotion detection task. The left plot of histograms in Figure 1 shows the $F1@2$ values on a run of *ESTeR* with the *EmoLex* lexicon on the different datasets. The coverage of words in Twitter vocabulary can be expected to be different from that of Wikipedia. We notice that combining information from both these resources yields better performance in 3 out of 5 datasets. A previous study by Klinger (2018), pointed out the domain similarity between *TEC* and *CrowdFlower* and the noise in *CrowdFlower* after a manual examination. Within *ESTeR*, both *TEC* and *CrowdFlower* benefit from using a focused corpus (*Twitter*) that is more reflective of their dataset domain.

Lexicon effect on *ESTeR*: In the middle plot of Figure 1, we show the effect of using the different lexicons on *ESTeR*. The $F1@2$ values achieved by *ESTeR* with *Wiki* matrix and the two lexicons *EmoLex* and *DepecheMood* are shown in this plot. While *EmoLex* is based on a general dictionary, the *DepecheMood* lexicon, uses vocabulary from news articles. Interestingly, the substantially smaller lexicon fares significantly better on all but one dataset (*TEC*). We attribute this effect to the quality of the lexicons. The *EmoLex* dictionary was created by asking annotators questions related to specific terms in the lexicon and them compiling them to reflect a binary association with an emotion (Mohammad and Turney, 2013). In contrast, the manual annotations obtained for news headlines were later converted to (word, emotion) association scores in *DepecheMood* (Araque et al., 2019). While this automatic process yields a large-scale dictionary, we note that within the *ESTeR* framework, having a smaller high-quality word associations seems to be more beneficial on average.

Performance of *ESTeR* variants: The rightmost plot in Figure 1 shows the performance of the three proposed variants *ESTeR*, *ESTeR:LexNorm*, *ESTeR:Lex2Sent* with *Combined* matrix and *EmoLex* lexicon on the five datasets. As described previously, the three variants have different interpretations: *ESTeR* is the probability that a random walker, starting at a randomly chosen word in a sentence, stops at *any* word in a lexicon whereas *ESTeR:LexNorm* penalizes large lexicons,

[6] All experiments were conducted on Xeon E5 2680 v2 2.80GHz with 64GB memory.

so that *every* lexicon word contributes equally. *ESTeR:Lex2Sent* is similar to *ESTeR:LexNorm*, but the walker moves from the lexicon to the sentence.

According to Figure 1, *ESTeR* outperforms the variants on 4 out of 5 datasets and is very close to the best performing variant for *CrowdFlower*. *ESTeR* and *ESTeR:LexNorm* result in a similar classification quality and outperform *ESTeR:Lex2Sent*. This is explainable; the walk from a relatively larger set of lexicon words quantifies the emotion association less precisely than the walk starting from a small set of sentence words. The lexicon normalization does not offer much benefit: a lexicon covers a range of words for a given emotion and it is restrictive to require the sentence to reflect all of them.

3.4 Comparison with baselines

Based on the experiments above, we choose *ESTeR* in combination with *EmoLex* lexicon and *Combined* matrix to compare against the baselines in Table 3. Additionally, we include results with the best-performing combination among (*ESTeR* variants, matrices and lexicons based on F1@2 scores) as *ESTeR**[7] entries. The best and second best performance for the two measures are highlighted in this table.

The first set of entries in this table uses state-of-the-art emotion-aware embeddings whereas the second set is based on word overlaps with the lexicon. Lexicon-based baselines are highly dependent on coverage of the words in the dataset and a given lexicon. Not surprisingly, this is reflected in the variation in performance with these baselines across the datasets. In comparison, the emotion-enriched embeddings are generated for capturing similarities and dissimilarities between words in a latent space. Hence, although emotions and sentences can be represented in embedding spaces, *ESTeR* is able to effectively harnesses word co-occurrence space to obtain a better performance on the classification task.

From Table 3, we observe that despite using a generic *EmoLex* lexicon and *Combined* graph, we still feature among the top-2 performing models for most datasets and outperform the baselines in most cases. Furthermore, by incorporating representative lexicons and matrices (the *ESTeR** entries), we obtain the best performance on all measures for three out of five datasets and the best F1@2 for all datasets. To summarize, *ESTeR* is able to effectively combine information from a general corpus and a focused word-association lexicon to provide a robust and competitive method for unsupervised emotion identification.

Method	F1-score		Jaccard	
	@1	@2	@1	@2
SemEval2018				
ewe-uni300	0.165	0.210	0.139	0.155
emo2vec100	**0.375**	0.379	0.259	0.254
sswe-u50	0.176	0.253	0.143	0.181
EmoLex	0.307	0.324	0.259	0.252
DepecheMood	0.287	0.401	0.237	*0.308*
ESTeR	*0.324*	*0.403*	**0.275**	0.305
ESTeR*	0.323	**0.430**	*0.265*	**0.324**
SSEC				
ewe-uni300	0.189	0.242	0.166	0.187
emo2vec100	*0.222*	0.240	0.148	0.156
sswe-u50	0.157	0.257	0.141	0.187
EmoLex	0.191	0.213	0.168	0.162
DepecheMood	0.202	*0.313*	0.180	*0.240*
ESTeR	0.209	0.277	*0.186*	0.209
ESTeR*	**0.228**	**0.325**	**0.205**	**0.248**
DENS				
ewe-uni300	0.128	0.175	0.128	0.131
emo2vec100	0.138	0.143	0.102	0.093
sswe-u50	0.066	0.157	0.066	0.117
EmoLex	**0.259**	*0.300*	**0.259**	*0.225*
DepecheMood	0.067	0.155	0.067	0.116
ESTeR	*0.241*	**0.301**	*0.241*	**0.226**
ESTeR*	*0.241*	**0.301**	*0.241*	**0.226**
TEC				
ewe-uni300	*0.309*	*0.397*	*0.309*	*0.298*
emo2vec100	0.298	0.291	0.208	0.192
sswe-u50	0.200	0.331	0.200	0.248
EmoLex	0.212	0.162	0.212	0.121
DepecheMood	0.150	0.218	0.150	0.163
ESTeR	0.306	0.324	0.306	0.243
ESTeR*	**0.398**	**0.421**	**0.398**	**0.316**
CrowdFlower				
ewe-uni300	0.296	*0.333*	0.296	*0.250*
emo2vec100	0.283	0.271	0.197	0.179
sswe-u50	0.164	*0.333*	0.164	*0.250*
EmoLex	0.193	0.126	0.193	0.094
DepecheMood	0.196	0.250	0.196	0.188
ESTeR	*0.323*	0.320	*0.323*	0.240
ESTeR*	**0.422**	**0.355**	**0.422**	**0.266**

Table 3: Comparison of classification quality with baselines in Section 2.5. *ESTeR* is run with *Combined* co-occurrence matrix and *EmoLex* lexicon. *ESTeR** denotes the best-performing combination of (lexicon, matrix, and *ESTeR* variant) choices. The **best** and *second-best* performances are highlighted.

4 Study of COVID-19 Tweets

We present an analysis of tweets related to the ongoing COVID-19 pandemic using *ESTeR* to high-

[7] *ESTeR** is a combination (*ESTeR:Lex2Sent*, *EmoLex*, *Combined*) for *SemEval2018* dataset; (*ESTeR:Lex2Sent*, *DepecheMood*, *Twitter*) for *SSEC*; (*ESTeR*, *EmoLex*, *Combined*) for *DENS*; (*ESTeR*, *EmoLex*, *Twitter*) for *TEC*; and (*ESTeR*, *EmoLex*, *Twitter*) for *CrowdFlower*.

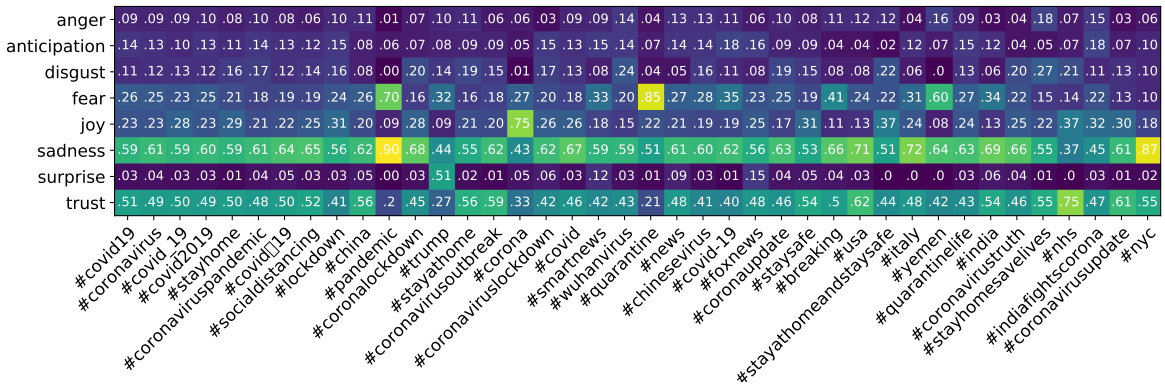

Figure 2: Top-40 frequent hashtags and the percentage of emotion labels assigned to the tweets with the hashtags.

light its usefulness in uncovering macro-level emotion trends despite zero labeled data. We study a random sample of 17K English tweets from the last week of March collected by Panacea labs[8]. This subset of tweets was analyzed using *ESTeR* method with *Combined* matrix and *EmoLex* lexicon. To obtain emotion labeling, we assign each tweet to 2 categories with the highest score. In the interest of space, we provide a summary of our findings in this section and provide further details in Appendix B for the interested reader.

In Figure 2 we show top frequent hashtags.[9] Each number in the matrix is the emotion frequency for the hashtag, i.e., the number of times the emotion is assigned to a tweet with the hashtag, divided by the total number of tweets with this hashtag in the dataset. We observe that, uniformly, *sadness* and *trust* are dominant emotions. General COVID-19 hashtags are mostly related to *sadness*, then to *trust* (due to the large volume of mentions of authorities and facts) and *fear*. Social tweets, which call to stay at home and embrace social distancing, also often have reassuring, comforting, and uplifting content, and thus are relatively often marked as *joy*. As expected, the majority of tweets with #pandemic tag get assigned to *sadness* and *fear*. Notably, the tag #trump gets most often assigned to *surprise* and infrequently to *trust* (probably reflective of public opinion due to his changing stance on COVID-19). Tags #quarantine and #yemen once again expectedly show a high assignment rate to *fear*. #NHS stands for United Kingdom National Health Service and is dominantly assigned to *trust*, *sadness*, and *joy* highlighting public emotion towards the struggles of the healthcare workers, as well as gratitude from the society.

[8] https://github.com/thepanacealab/covid19_twitter

[9] Further analysis is included in Appendix B.

5 Related Work

As part of affective computing, various research communities are studying emotion identification models via gestures and facial expressions (Barros et al., 2015), voice (Mitsuyoshi et al., 2017) as well as user-generated text (Canales and Martínez-Barco, 2014; Aguilar et al., 2019). In particular, text-based emotion detection and mood analysis has attracted significant research focus through task challenges (Mohammad et al., 2018; Hsu and Ku, 2018) due to the abundance of user-generated content on social media and microblogging platforms that captures public mood on various events in social, political, and economic spheres (Bollen et al., 2011a; Nguyen et al., 2014; Khanpour and Caragea, 2018).

Emotion detection was studied in various settings including social media content (Preoţiuc-Pietro et al., 2016), literature (Liu et al., 2019), TV-show transcripts (Zahiri and Choi, 2018) and conversations (Majumder et al., 2019) using supervised learning approaches. The best performing models are based on deep learning with labeled data and other knowledge resources such as lexicons and word embeddings (Abdul-Mageed and Ungar, 2017; Zhong et al., 2019; Islam et al., 2019). Transfer learning and multi-task learning techniques were also studied for reducing labeled data requirements for supervision (Zhang et al., 2018; Tafreshi and Diab, 2018; Dankers et al., 2019). Previous studies include those on automatic and crowd-sourced building of lexicons (Mohammad and Turney, 2013; Araque et al., 2019; Rao et al., 2014; Buechel and Hahn, 2018) as well as learning emotion-enhanced word embeddings (Agrawal et al., 2018; Xu et al., 2018; Saravia et al., 2018).

6 Concluding Remarks

We proposed a random walk based model for unsupervised emotion detection in text using word associations from emotion lexicons and word co-occurrences from a general corpus. Our solution efficiently computes emotion scores at a dataset level as well as provides a probabilistic interpretation of scores. We showed superior performance of our model over existing unsupervised baselines on several recent, real-world datasets. In future, we would like to study other graph-based scoring functions to further improve performance (Boudin, 2013). In particular, we are interested in minimally-supervised representations that can apply to a range of related tasks that involve emotions such as sarcasm, stress and insult detection, abusive language classification, and personality recognition (Xu et al., 2018).

Acknowledgments

This research/project is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG-GC-2019-001). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

References

Muhammad Abdul-Mageed and Lyle Ungar. 2017. Emonet: Fine-grained emotion detection with gated recurrent neural networks. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 718–728.

Ameeta Agrawal, Aijun An, and Manos Papagelis. 2018. Learning emotion-enriched word representations. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 950–961.

Gustavo Aguilar, Viktor Rozgic, Weiran Wang, and Chao Wang. 2019. Multimodal and multi-view models for emotion recognition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 991–1002.

Mario Ezra Aragón, Adrian Pastor López-Monroy, Luis Carlos González-Gurrola, and Manuel Montes-y Gómez. 2019. Detecting depression in social media using fine-grained emotions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1481–1486.

Oscar Araque, Lorenzo Gatti, Jacopo Staiano, and Marco Guerini. 2019. Depechemood++: a bilingual emotion lexicon built through simple yet powerful techniques. *IEEE transactions on affective computing*.

Arvind Arasu, Jasmine Novak, Andrew Tomkins, and John Tomlin. 2002. Pagerank computation and the structure of the web: Experiments and algorithms. In *Proceedings of the Eleventh International World Wide Web Conference, Poster Track*, pages 107–117.

Pablo Barros, Doreen Jirak, Cornelius Weber, and Stefan Wermter. 2015. Multimodal emotional state recognition using sequence-dependent deep hierarchical features. *Neural Networks*, 72:140 – 151. Neurobiologically Inspired Robotics: Enhanced Autonomy through Neuromorphic Cognition.

Johan Bollen, Huina Mao, and Alberto Pepe. 2011a. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *ICWSM*.

Johan Bollen, Huina Mao, and Xiaojun Zeng. 2011b. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1 – 8.

Florian Boudin. 2013. A comparison of centrality measures for graph-based keyphrase extraction. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 834–838, Nagoya, Japan. Asian Federation of Natural Language Processing.

Sven Buechel and Udo Hahn. 2018. Emotion representation mapping for automatic lexicon construction (mostly) performs on human level. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2892–2904.

Lea Canales and Patricio Martínez-Barco. 2014. Emotion detection from text: A survey. In *Proceedings of the Workshop on Natural Language Processing in the 5th Information Systems Research Working Days (JISIC)*, pages 37–43.

Ankush Chatterjee, Umang Gupta, Manoj Kumar Chinnakotla, Radhakrishnan Srikanth, Michel Galley, and Puneet Agrawal. 2019. Understanding emotions in text using deep learning and big data. *Computers in Human Behavior*, 93:309–317.

Guan-Bin Chen and Hung-Yu Kao. 2015. Word co-occurrence augmented topic model in short text. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 20, Number 2, December 2015 - Special Issue on Selected Papers from ROCLING XXVII*.

Verna Dankers, Marek Rei, Martha Lewis, and Ekaterina Shutova. 2019. Modelling the interplay of metaphor and emotion through multitask learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2218–2229.

Mandar Deshpande and Vignesh Rao. 2017. Depression detection using emotion artificial intelligence. In *2017 International Conference on Intelligent Sustainable Systems (ICISS)*, pages 858–862. IEEE.

Jianyong Duan, Jiayuan Cui, Mingli Wu, and Hao Wang. 2018. Capturing semantic similarity for words in wikipedia with random walk. In *2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS)*, pages 709–713. IEEE.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Paul Ekman. 2016. What scientists who study emotion agree about. *Perspectives on Psychological Science*, 11(1):31–34.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009.

Taher Haveliwala, Sepandar Kamvar, and Glen Jeh. 2003. An analytical comparison of approaches to personalizing pagerank. Technical report, Stanford.

Taher H Haveliwala. 2003. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *IEEE transactions on knowledge and data engineering*, 15(4):784–796.

Chao-Chun Hsu and Lun-Wei Ku. 2018. SocialNLP 2018 EmotionX challenge overview: Recognizing emotions in dialogues. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*, pages 27–31.

Tianran Hu, Anbang Xu, Zhe Liu, Quanzeng You, Yufan Guo, Vibha Sinha, Jiebo Luo, and Rama Akkiraju. 2018. Touch your heart: A tone-aware chatbot for customer care on social media. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 1–12.

Jumayel Islam, Robert E. Mercer, and Lu Xiao. 2019. Multi-channel convolutional neural network for twitter emotion and sentiment recognition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1355–1365.

Hamed Khanpour and Cornelia Caragea. 2018. Fine-grained emotion detection in health-related online posts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1160–1166.

Sunghwan Mac Kim, Alessandro Valitutti, and Rafael A Calvo. 2010. Evaluation of unsupervised emotion models to textual affect recognition. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 62–70. Association for Computational Linguistics.

Roman Klinger et al. 2018. An analysis of annotated corpora for emotion classification in text. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2104–2119.

Jing Kong, Alex Scott, and Georg M. Goerg. 2016. Improving semantic topic clustering for search queries with word co-occurrence and bigraph co-clustering.

Maximilian Köper, Evgeny Kim, and Roman Klinger. 2017. Ims at emoint-2017: emotion intensity prediction with affective norms, automatically extended resources and deep learning. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 50–57.

Chen Liu, Muhammad Osama, and Anderson De Andrade. 2019. Dens: A dataset for multi-class emotion analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6294–6299.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander F. Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive RNN for emotion detection in conversations. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI*, pages 6818–6825.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411.

Shunji Mitsuyoshi, Mitsuteru Nakamura, Yasuhiro Omiya, Shuji Shinohara, Naoki Hagiwara, and Shinichi Tokuno. 2017. Mental status assessment of disaster relief personnel by vocal affect display based on voice emotion recognition. *Disaster and Military Medicine*.

Saif M Mohammad. 2012. # emotional tweets. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 246–255. Association for Computational Linguistics.

Saif M. Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 Task 1: Affect in tweets. In *Proceedings of International Workshop on Semantic Evaluation (SemEval-2018)*, New Orleans, LA, USA.

Saif M Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2017. Stance and sentiment in tweets. *ACM Transactions on Internet Technology (TOIT)*, 17(3):1–23.

Saif M Mohammad and Peter D Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Shreshtha Mundra, Anirban Sen, Manjira Sinha, Sandya Mannarswamy, Sandipan Dandapat, and Shourya Roy. 2017. Fine-grained emotion detection in contact center chat utterances. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 337–349. Springer.

Arch W Naylor and George R Sell. 2000. *Linear operator theory in engineering and science*. Springer Science & Business Media.

Venkata K. Neppalli, Cornelia Caragea, Anna Squicciarini, Andrea Tapia, and Sam Stehle. 2017. Sentiment analysis during hurricane sandy in emergency response. *International Journal of Disaster Risk Reduction*, 21:213 – 222.

Thin Nguyen, Dinh Phung, Brett Adams, and Svetha Venkatesh. 2014. Mood sensing from social media texts and its applications. In *Knowledge and Information Systems*, volume 39, pages 667–702.

Endang Wahyu Pamungkas. 2019. Emotionally-aware chatbots: A survey. *CoRR*, abs/1906.09774.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Robert Plutchik. 2001. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist*, 89(4):344–350.

Marco Polignano, Pierpaolo Basile, Marco de Gemmis, and Giovanni Semeraro. 2019. A comparison of word-embeddings in emotion detection from text using bilstm, cnn and self-attention. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*, pages 63–68.

Daniel Preoţiuc-Pietro, H. Andrew Schwartz, Gregory Park, Johannes Eichstaedt, Margaret Kern, Lyle Ungar, and Elisabeth Shulman. 2016. Modelling valence and arousal in Facebook posts. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 9–15.

Y Rao, Q Li, L Wenyin, Q Wu, and Quan X. 2014. Affective topic model for social emotion detection. In *Neural Networks*, volume 58.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. CARER: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697.

Julio Savigny and Ayu Purwarianti. 2017. Emotion classification on youtube comments using word embedding. In *2017 International Conference on Advanced Informatics, Concepts, Theory, and Applications (ICAICTA)*, pages 1–5. IEEE.

Hendrik Schuff, Jeremy Barnes, Julian Mohme, Sebastian Padó, and Roman Klinger. 2017. Annotation, modelling and analysis of fine-grained emotions on a stance and sentiment detection corpus. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 13–23.

Armin Seyeditabari, Narges Tabari, and Wlodek Zadrozny. 2018. Emotion detection in text: a review. *arXiv preprint arXiv:1806.00674*.

Armin Seyeditabari and Wlodek Zadrozny. 2017. Can word embeddings help find latent emotions in text? preliminary results. In *The Thirtieth International Flairs Conference*.

Tiffany Watt Smith. 2015. *The book of human emotions: An encyclopedia of feeling from anger to wanderlust*. Profile Books.

Jacopo Staiano and Marco Guerini. 2014. 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)*, pages 427–433.

Shabnam Tafreshi and Mona Diab. 2018. Emotion detection and classification in a multigenre corpus with joint multi-task deep learning. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2905–2913.

Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. 2014. 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)*, pages 1555–1565.

Wei Wei, Jiayi Liu, Xianling Mao, Guibing Guo, Feida Zhu, Pan Zhou, and Yuchong Hu. 2019. Emotion-aware chat machine: Automatic emotional response generation for human-like emotional interaction. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 1401–1410.

Peng Xu, Andrea Madotto, Chien-Sheng Wu, Ji Ho Park, and Pascale Fung. 2018. Emo2vec: Learning generalized emotion representation by multi-task training. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 292–298.

Ruifeng Xu, Tao Chen, Yunqing Xia, Qin Lu, Bin Liu, and Xuan Wang. 2015. Word embedding composition for data imbalances in sentiment and emotion classification. *Cognitive Computation*, 7(2):226–240.

Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. 2013. A biterm topic model for short texts. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1445–1456.

Majid Yazdani and Andrei Popescu-Belis. 2010. A random walk framework to compute textual semantic similarity: a unified model for three benchmark tasks. In *2010 IEEE Fourth International Conference on Semantic Computing*, pages 424–429. IEEE.

Chih-Kuan Yeh, Wei-Chieh Wu, Wei-Jen Ko, and Yu-Chiang Frank Wang. 2017. Learning deep latent spaces for multi-label classification. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 2838–2844. AAAI Press.

Sayyed M Zahiri and Jinho D Choi. 2018. Emotion detection on tv show transcripts with sequence-based convolutional neural networks. In *Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence*.

Yuxiang Zhang, Jiamei Fu, Dongyu She, Ying Zhang, Senzhang Wang, and Jufeng Yang. 2018. Text emotion distribution learning via multi-task convolutional neural network. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4595–4601. International Joint Conferences on Artificial Intelligence Organization.

Peixiang Zhong, Di Wang, and Chunyan Miao. 2019. Knowledge-enriched transformer for emotion detection in textual conversations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 165–176.

Zahari Zlatev. 1991. *Computational methods for general sparse matrices*, volume 65. Springer Science & Business Media.

A Further Details on Experimental Settings

A.1 Datasets

We provide more details on our datasets:

The *SemEval2018* (Mohammad et al., 2018) dataset is manually annotated with 8 Plutchek's prime emotions and three more: *love*, *optimism*, *pessimism*. We keep only 8 prime emotions in the annotation. The total size is 10.5K, multi-labeling is allowed, the maximum number of labels per a tweet is 6 and median is 2.

For the *SSEC* (Schuff et al., 2017) dataset, the annotation is done using 8 Plutchek's prime emotions but the dataset is available is several versions, we used "0.5" version, where each label is voted for by more than a half of annotators. The total size is 3.3K. The maximum number of labels per a tweet is 5 and median is 1.

In *DENS* (Liu et al., 2019), the passages are manually annotated using 8 Plutchek's emotions plus *neutral*, with *trust* is substituted by *love* since the labelers could recognize *trust* better in romantic context. According to Plutchek, *love* is a combinations of *trust* and *joy*). We substitute *love* back with *trust*. Only the majority-voted annotations are preserved. The number of passages, annotated with 8 Plutchek's emotions is 8K. The maximum number of labels per a passage is 2, median is 1.

TEC (Mohammad, 2012) is manually annotated using six Ekman's emotions. The total size is 20.5K. No multi-labeling is allowed.

For *CrowdFlower*[10], we used the mapping proposed (Klinger et al., 2018) to obtain labels from 8 Plutchek's emotions from their 13 non-standard emotional categories, followed by a majority-based annotation aggregation. The total size is 31.2K. The maximum number of labels per a tweet is 2 and median is 1.

We processed all datasets using NLTK TweetTokenizer[11]. Emoticons are preserved and only non-stopwords with lengths greater than one character are retained.

A.2 Co-occurrence matrices

Wiki co-occurrence matrix was obtained from Wikipedia collected in Feb 2020 and has approximately 5.2M documents. We apply term frequency, document frequency thresholds of 100 and 5 for collecting the term dictionary and only keep edges between words that occur within a window of 5 and with edge frequency threshold of 200. The final co-occurrence matrix contains 39K words and 1.7M non-zero entries. *Twitter* co-occurrence matrix was obtained from Twitter dataset *sentiment140* (Go et al., 2009) with 1.6M general tweets. Two words are considered co-occurring if they appear in the same tweet. The co-occurrence threshold is set to 10. The co-occurrence matrix contains 17.7K tokens and 1.3M non-zero entries. *Combined* is the (unweighted) combination of the two matrices *Wiki* and *Twitter* co-occurrence matrices with 47.7K tokens and 1.9M non-zero entries.

[10] https://data.world/crowdflower/sentiment-analysis-in-text

[11] https://www.nltk.org/api/nltk.tokenize.html

B Detailed Case Study Findings

We continue the case study of *COVID19* dataset by *ESTeR* with *Combined* matrix and *EmoLex* lexicon. Recall that to obtain emotion labeling of the tweets, we assign each tweet to (at least, in case of ties) 2 categories with the highest score.

In Table 4 we group hashtags by top-2 emotion labels, which are most frequently assigned to the tweets with the corresponding hashtags. Note that the order of the emotion labels matters. For example, group 1 has *sadness* as the most frequent label and *trust* as the second most frequent label; $(trust, sadness)$ produces a different cluster. To generate Table 4 we go through the most frequent (most popular) hashtags in the descending order. Each hashtag is added to a cluster based on top-2 emotions most frequently assigned for tweets with this hashtag. The clusters are ordered based on the maximum popularity of the hashtags they contain. We report at most 5 the most popular hashtags of each cluster. Inside each cluster the hashtags are sorted by the popularity. Table 4 shows top-6 clusters. Since none of them have *anger*, *disgust*, or *anticipation* as the most frequent emotion, we report also clusters number 9, 12, and 22 - the clusters with the highest rank having one of these missing emotions as the most frequently associated.

Interestingly, $(anticipation, sadness)$ cluster covers #stocks hashtag. $(Disgust, surpirse)$ covers American political hashtags as well as mentions of pop artists. $(Anger, surprise)$ covers business-related news. Unlike health-related topics, people tend to express less empathy and more discontent with COVID-19 impact on the economy. Cluster of $(joy, sadness)$ includes tags #love as well as country names, as these tags are often used in tweets

N	Emotions	Hashtags
1	sadness, trust	#covid19, #coronavirus, #covid_19, #covid2019, #stayhome, #coronaviruspandemic, #covid-19, #socialdistancing, #lockdown, #china
2	sadness, fear	#pandemic, #yemen, #maga, #chinesevirus19, #hantavirus, #fakenews, #coronavirusoubreak, #covid2019india, #potus, #whencoronavirusisover
3	surprise, sadness	#trump, #market, #marijuana, #f1, #mobility, #microsoft, #hilarious, #strike, #awareness, #executive
4	trust, sadness	#stayathome, #staysafe, #stayhomesavelives, #nhs, #indiafightscorona, #coronavirusupdate, #21daylockdown, #stayhomestaysafe, #rt, #21dayslockdown
5	joy, sadness	#corona, #love, #spain, #paris, #artist, #frankfurt, #radio, #gaza, #karnataka, #independent
6	fear, sadness	#quarantine, #government, #live, #daca, #brexit, #google, #recession, #stimulus, #mnd, #dementia
9	anger, surprise	#business, #jimin, #smallbiz, #calm, #minnesota
12	disgust, surprise	#trumpliesamericansdie, #moron, #followtrick, #cardib
22	anticipation, sadness	#stocks, #cruise, #herdimmunity, #daily, #slovakia, #stayingathome

Table 4: Popular hashtags, grouped by the emotions, which are the most frequently assigned to the corresponding tweets.

Emotion	Tweets
anger	Indeed!China cannot be trusted... MUST NOT be trusted!#Wuhan #nCov outbreak is an threat to the world!It is... Because #Democrats try to sneak in crap like paying Illegal aliens or New Green Deal. Pay the People, keep the... Do You Need to Rise in Business? Knock Me Now. #bitcoin #jungkook #SundayFunday #promote #stayhome...
anticipation	”In this clip he1. Denies WHO’s coronavirus death rate based on “hunch””2. Calls #coronavirus ””corona flu””... Ingratitude: Top Italian newspaper calls Russian #Covid19 aid ’useless’, implies Putin using medical mission... For just #100,000We come to ur house dressed as #covid19 rescue team to rescue u from ur wife, then take u to...
disgust	News Oz: Politicians Are Not Letting the Coronavirus Crisis Go to Waste #newsoz.org #news Commentary In... Boom. You’re nuts. You and The Trump Shit Show should make like COVID19 and #GoAway CNN didn’t loo... John Oliver Unloads On Right Wing Media’s ‘Death Cult’ Over Coronavirus — HuffPost Canada#MAGA2020...
fear	#Cholera (1899-1924 ˜23 yrs) 6th pandemic of Cholera bacteria infection of Europe, Asia and Africa with death... Indeed!China cannot be trusted... MUST NOT be trusted!#Wuhan #nCov outbreak is an threat to the world!It is... I’m very wary of people coming from Abuja and Lagos. As far as I’m concerned they’re all vectors. No joke...
joy	Official Isolation Day 1Stay safe,stay happy and trust God.#StaySafeNigeria #COVID19 #churchboy #pastorson... ”Neil Diamond’s #CoronaVirus Version of Sweet Caroline: ””Hands... Washing hands... Don’t touch me... I... Good morning beautiful world. Sending out positive healthy vibes to everyone across the globe. Stay safe &...
sadness	Yuan Shun “Red &; Black,10000 B.C-2028 A.D#59, 2020. Ink and colour on paper.#architecture #art #artist... I think this is mother earth’s way of telling us to sit our arses DOWN while she fixes the fecking mess we’ve... .@rocklandgov Exec Ed Day says #Rockland already hit hard by bottoming out of sales tax revenue. But he...
surprise	Why is Chinese gov backed business #GreenlandGroup allowed to bulk buy urgent medical supplies hazmat... Do You Need to Rise in Business? Knock Me Now. #bitcoin #jungkook #SundayFunday #promote #stayhome... #COVID19 UAE: TRA unblocks Skype for Business, Google Hangouts amid COVID-19 outbreak also Micro...
trust	Official Isolation Day 1Stay safe,stay happy and trust God.#StaySafeNigeria #COVID19 #churchboy #pastorson... #DevelopedEconomies #calls #Covid19 WTO chief sees sharp fall in trade, calls for global solutions to COVID... Ingratitude: Top Italian newspaper calls Russian #Covid19 aid ’useless’, implies Putin using medical mission...

Table 5: Tweets with the highest association *ESTeR* score to an emotion from *COVID19*.

of sympathy. $(Trust, sadness)$ cluster consists of tweets supporting social measures, expressing sympathy for health workers, and generally uniting tweets. $(Surprise, sadness)$ is related to US politics and market news. $(Sadness, fear)$ covers not only COVID-19-related tags, but also Yemen armed conflict, fakenews warning. The top cluster $(sadness, trust)$: is general coronavirus tags, *trust* has a high presence due to a lot of comments on official information.

In Table 5, for each emotion category $e \in \mathcal{E}$ we report top 3 tweets with the highest association *ESTeR* score to e. To present a constructive examples, we consider tweets with at least 15 tokens. Due to the nature of the dataset, most of the tweets express emotions such as *fear* and *sadness*. However, there are still tweets labeled as *joy*, which contain jokes or express hope and optimism. Interestingly, *disgust* label brings up comments on political news in the time of pandemic. *Surprise* is represented with tweets with a question, *trust* labels tweets with official mentions and economics-related news.