Systematic Evaluation of a Framework for Unsupervised Emotion Recognition for Narrative Text

Samira Zad & Mark A. Finlayson
School of Computing and Information Sciences
Florida International University
11200 SW 8th St., Miami, FL 33199, USA
{szad001,markaf}@fiu.edu

Abstract

Identifying emotions as expressed in text (a.k.a. text emotion recognition) has received a lot of attention over the past decade. Narratives often involve a great deal of emotional expression, and so emotion recognition on narrative text is of great interest to computational approaches to narrative understanding. Prior work by Kim et al. (2010) was the work with the highest reported emotion detection performance, on a corpus of fairy tales texts. Close inspection of that work, however, revealed significant reproducibility problems, and we were unable to reimplement Kim's approach as described. As a consequence, we implemented a framework inspired by Kim's approach, where we carefully evaluated the major design choices. We identify the highest-performing combination, which outperforms Kim's reported performance by 7.6 F_1 points on average. Close inspection of the annotated data revealed numerous missing and incorrect emotion terms in the relevant lexicon, WordNetAffect (WNA; Strapparava and Valitutti, 2004), which allowed us to augment it in a useful way. More generally, this showed that numerous clearly emotive words and phrases are missing from WNA, which suggests that effort invested in augmenting or refining emotion ontologies could be useful for improving the performance of emotion recognition systems. We release our code and data to definitely enable future reproducibility of this work.

1 Introduction

Emotion is a primary aspect of communication, and can be transmitted across many modalities including gesture, facial expressions, speech, and text. Because of this importance, automatic emotion recognition is useful for many applications, including for automated narrative understanding. A narrative is "a representation of connected events and characters that has an identifiable structure, is bounded in space and time, and contains implicit or explicit messages about the topic being addressed" (Kreuter et al., 2007, p. 222), and narratives are often used to express the emotions of authors and characters, as well as induce emotions in audiences. For many narratives—one need only consider romances such as *Romeo and Juliet* or the movie *Titanic*—it is no exaggeration to say that lacking an understanding of emotion leads to a seriously impoverished view of the meaning of the narrative.

Emotion recognition is a challenging problem on account of the complex relationship between felt emotion and linguistic expression. This includes not only standard natural language processing challenges, such as polysemous words and the difficulty of coreference resolution (Uzuner et al., 2012; Peng et al., 2019), but also emotion-specific challenges such as how context can subtly change emotional interpretations (Cowie et al., 2005). These technical challenges are exacerbated by a shortage of quality labeled data addressing this task.

There has been much prior work on emotion recognition. With regard to narrative specifically, Kim et al. (2010) reported a high-performing approach to emotion recognition on a corpus of fairy tales texts (Alm, 2008). This approach involved an unsupervised learning framework for emotion recognition in textual data, using a modified form of Ekman's psychological theory of emotion (joy, anger, fear, sadness; Ekman, 1992b). In that work, they used the WordNetAffect (WNA) and ANEW (Affective Norm for English Words) emotion lexicons to construct a semantic space. Each sentence is placed in the space using *tf-idf* weights for emotion words found in the lexicons. They then tested three methods—Non-negative Matrix Factorization (NMF), Latent Semantic Analysis (LSA), probablistic Latent Semantic Analysis (pLSA)—for compressing the space to extract features of the constructed vector space model, reduce noise,

Proceedings of the 1st Joint Workshop on Narrative Understanding, Storylines, and Events, pages 26–37
July 9, 2020.

and eliminate outliers. Finally, the framework used cosine-similarity to label sentences by evaluating how similar they are compared to standard vectors generated based on WNA entries strongly associated with emotion lexicon (more specifically an extension of WNA). The best performing method was NMF, which they reported achieved an average emotion recognition F_1 of 0.733.

Close inspection of the work, however, revealed significant reproducibility problems. Despite our best efforts we were unable to reproduce results anywhere near Kim's reported performance; indeed, our best attempt yielded only roughly 0.25 F_1. This was due to several reasons. First, the paper lacked information on model hyper-parameters. Second, the paper omitted descriptions of key NMF steps, including how to identify representative features and what features should be removed before semantic space compression. Third, the paper did not explain how to adapt NMF to deal with the sparse matrices that occur in textual NMF models. Fourth, certain resources associated with WNA either were not correctly identified, or are no longer available. These omissions prevented us from reproducing their models to any degree of accuracy.

Therefore, we undertook to do a systematic exploration of the design space described in Kim et al. (2010). We examined the highest performing vector space compression techniques reported by Kim *et al.* (NMF), as well as Principle Component Analysis (PCA) and Latent Dirchelet Allocation (LDA) which were reported as high-performing techniques in other work. We show that NMF indeed performs the best, and we clearly explain our experimental setup including methods for identifying relevant features and handling sparse text matrices. The PCA and NMF methods implemented in this paper are based on the works of Mairal et al. (2009) and Boutsidis and Gallopoulos (2008) respectively which have implemented mechanisms that works for a large sparse matrix (in our case, $1,090 \times 2,405$). This work resulted in an improvement of performance of roughly 7.6 points of F_1 over Kim's reported results. We release our code and data to facilitate future work[1].

The rest of this paper is structured as follows. We briefly review psychological models of emotions, describe several key emotion language resources, and outline a number of well-known emotion recognition models (§2). We then describe our adapted unsupervised emotion recognition method, giving detailed descriptions of all steps, parameters, and resources needed (§3). We next describe the performance of our method on Alm's corpus of fairy tales (Alm, 2008), which was annotated for emotion on a per-sentence level (§4). Finally, we identify some unsolved challenges that point toward future work (§5), and summarize our contributions (§6).

2 Related Work

2.1 Psychological emotion theories

Theories of emotion go back to the ancient Greeks and Romans, and have been a recurring theme of inquiries into the nature of the human experience throughout history, including famous proposals by Charles Darwin and William James in the 19th century (Darwin and Prodger, 1998; James, 1890). Modern psychological theories of emotion may be grouped into two types: *categorical* and *dimensional* (Calvo and Mac Kim, 2013). Categorical psychological models propose discrete basic emotions, e.g., Oatley and Johnson-Laird's (1987) with five basic emotions, several models with six basic emotions (Ekman, 1992b; Shaver et al., 1987), Parrott's model of six basic emotions arranged in a three-level tree (2001), Panksepp's model with seven emotions (1998), and Izard's with ten (2007).

Dimensional psychological models, by contrast, determine emotions by locating them in a space of dimensions (usually two to four) that might include arousal, valence, intensity, etc. These include two dimensional models such as Russell's circumplex model (1980), Scherer's augmented circumplex (2005), and Whissell's model (Cambria, 2016). Lövheim's model (2012) is an example that uses three dimensions, while Ortony et al. (1990), Fontaine et al. (2007), and Cambria et al. (2012) proposed four-dimensional models.

Finally, there are also models which combine both categorical and dimensional aspects, called *hybrid* models, the most prominent of which is Plutchik's wheel and cone model with eight basic emotions (Plutchik, 1980, 1984, 2001).

Of all the many emotion models that have been proposed, Ekman's 6 category model (anger, disgust, fear, happiness, sadness, surprise) is by far the most popular in computational approaches, partly because of its simplicity, and partly because it has been successfully applied to automatic facial emo-

[1] Code and data may be downloaded from https://doi.org/10.34703/gzx1-9v95/03RERQ

tion recognition (Zhang et al., 2018; Suttles and Ide, 2013; Ekman, 1992b,a, 1993). This is despite that some researchers have doubts that Ekman's model is complete, as it seems to embed a Western cultural bias (Langroudi et al., 2018). In our own review of emotion recognition systems, as discussed below, the highest performing system reported for narrative text was described by Kim et al. (2010). In that work, they used a four-label subset of Ekman's model (happiness, anger, fear, and sadness), and this is the model we adopt in this paper.

2.2 Emotion Lexicons

One of the key language resources for emotion recognition in text is an emotion lexicon, which is simply a list of words associated with emotion categories. Emotion lexicons can be used both in rule-based and machine-learning-based recognition methods. There are two types of emotion lexicons. One is general purpose emotion lexicons (GPELs) which specify the generic sense of emotional words. GPELs sometimes express emotions as a score, and can be applied to any domains. Prominent GPELs include WordNet Affect (WNA; Strapparava and Valitutti, 2004), the Wisconsin Perceptual Attribute Rating Database (WPARD; Medler et al., 2005), Linguistic Inquiry and Word Count (LIWC; Pennebaker et al., 2001), and the National Research Council (NRC) and NRC Hashtag lexicons (Mohammad and Turney, 2010; Mohammad et al., 2013). The second type of lexicon are domain specific emotion lexicons (DSELs) which are targeted at specific domains for emotion recognition. Bandhakavi et al. (2014), for example, proposes a domain-specific lexicon for emotional tweets. Table 1 compares the details of several key GPELs.

WordNet Affect Version 1.1 Kim et al. used WordNet Affect (WNA; Strapparava and Valitutti, 2004), which builds upon the general WordNet database (Fellbaum, 1998). WNA classifies 280 WordNet *Noun* synsets into an emotion hierarchy rooted in an augmented version of Ekman's basic emotions, and partially depicted in Figure 1. WordNet links an additional 1,191 *Verb*, *Adverb*, and *Adjective* synsets to this core *Noun*-focused hierarchy. These synsets represent approximately 3,500 English lemma-POS pairs.

2.3 Emotion Recognition Approaches

There have been at least one hundred papers describing approaches to emotion recognition in text

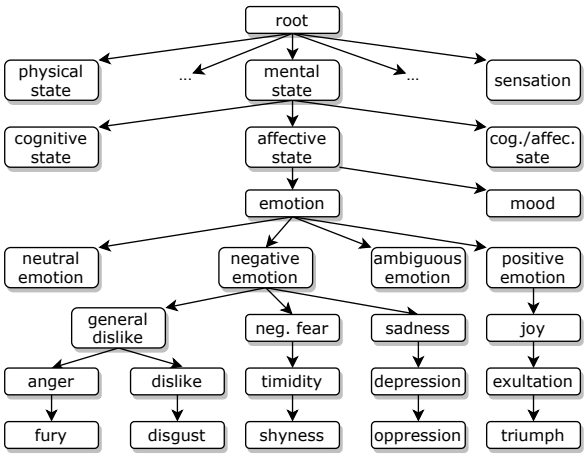


Figure 1: Hierarchy of emotions in WordNet Affect Version 1.1.

(Calefato et al., 2017; Teng et al., 2007; Shaheen et al., 2014). Here we review a selection of approaches that have been applied to narrative-like or narrative-related discourse types. It is important to remember that all of these approaches use different data and different theories, often involving different numbers of labels. All things being equal, classification results usually degrade as the number labels increases; therefore the performance of each system can only be loosely compared.

Strapparava and Mihalcea (2008) described a system for recognizing emotions in news headlines. They extracted 1,250 news headlines from a variety of news websites (such as Google news, CNN, and online newspapers) and annotated them using Ekman's model—anger, disgust, fear, joy, sadness and surprise—splitting the data into a training set of 250 and a test set of 1,000 (this is called the *SemEval-2007* dataset). They tested five approaches: WNA-PRESENCE, LSA-SINGLE-WORD, LSA-EMOTION-SYNSET, LSA-ALL-EMOTION-WORDS, and NAIVEBAYES-TRAINED-ON-BLOGS. WNA-PRESENCE, which looked for headline words listed in WNA, provided the best precision at 0.38. The LSA-ALL-EMOTION-WORDS, which calculated the vector similarity between the six affect words and the LSA representation of the headline, led to the highest recall and F_1, at 0.90 and 0.176, respectively.

Aman and Szpakowicz (2008) used a Support Vector Machine (SVM) trained and tested on blog data for recognition Ekman's emotion classes, plus two additional classes: *mixed emotion*, and *no emotion*. Four human judges manually annotated 1,890 sentences from automatically retrieved blogs to cre-

Emotion Lexicons	Citation	Set of Emotions	Entries
WNA	Strapparava and Valitutti (2004)	A hierarchy of emotions	915 synsets
NRC / Emolex	Mohammad and Turney (2010)	Plutchik basic model 1980, neg./pos.	14,182
LIWC	Pennebaker et al. (2001)	Affective or not, neg./pos. anxiety, anger, sadness	5,690
NRC Hashtag	Mohammad et al. (2013)	Plutchik's basic model	32,400
WPARD	Medler et al. (2005)	Positive or negative	1,402
ANEW	Bradley and Lang (1999)	3D (valence, arousal,dominance)	1,035

Table 1: Emotion-related lexicons table. WNA= WordNet Affect; NRC= National Research Council in Canada; LIWC= Linguistic Inquiry and Word Count; WPARD= Wisconsin Perceptual Attribute Rating Database; ANEW= Affective Norms of English Words

ate the corpus. The features for the SVM were the presence of emotion words listed in Roget's thesaurus and WNA. F_1 measures for each emotion class ranged between 0.493 to 0.751, in each case surpass the baseline performance.

Tokuhisa et al. (2008) described a lexicon-based emotion recognition system for Japanese. They handcrafted emotion lexicon by identifying 349 emotion words from the Japanese Expression Evaluation (JEE) Dictionary classified into 10 different emotions: 3 positive (happiness, pleasantness, relief) and 7 negative (fear, sadness, disappointment, unpleasantness, loneliness, anxiety, and anger). They then used this lexicon to automatically assemble a labeled corpus of 1.3M emotion-provoking (EP) "events" (defined as a subordinate clauses which modifies an emotional statement). They then demonstrated a two-step method for emotion recognition, starting with SVM-based coarse sentiment polarity classification (positive, negative, or neutral) followed by kNN-based classification of non-neutral instances into the appropriate fine-grained emotion classes (3 for positive, 7 for negative). Their reported accuracies of between 0.5 and 0.8 for their best performing model.

Cherry et al. (2012) presented two supervised machine learning models for emotion recognition in suicide note sentences. They used the 2011 i2b2 NLP Challenge Task 2, which comprised 4,241 sentences in the training set, and 1,883 sentences in the test set, which were manually annotated with 13 emotion labels. A one-classifier-per-emotion approach yielded an F_1 of 0.55, while a latent sequence model that applied multiple emotion labels per sentence achieved an F_1 of 0.53. They noted that more than 73% of their training data lacked labels which limited the effectiveness of the training.

Bandhakavi et al. (2017) experimented with unigram mixture models (UMMs) for recognizing emotions in tweets, incident reports, news headlines, and blogs. Each corpus was manually annotated with different emotion theories: 280,000 tweets with Parrott's six primary emotions (Parrott, 2001), 1,250 news headlines and 5,500 blogs with Ekman's six emotion set, 7000 incident reports from the ISEAR dataset[2] labeled with a seven emotion set. One goal of the study was to compare the utility of domain-specific emotion lexicons with general purpose emotion lexicons (DSELs vs GPELs). They found that combining DSEL lexicon words with n-grams, part of speech tags, and additional words from sentiment lexicons yielded the highest performance of 0.60 F_1 on the blog data.

Kim et al. (2010) reported the highest performing emotion recognition system on narrative text. Among their data was a set of 176 fairy tales whose 15,087 sentences were labeled by Alm (2008) with a four-emotion subset of Ekman's theory (anger, fear, joy, and sadness). They demonstrated an unsupervised approach, where each sentence is transformed into a vector in a space of emotion words (drawn from WNA and ANEW), and then compressed using a dimension reduction technique (NMF, LSA, or pLSA). These vectors were then compared to reference vectors in the same space that were computed for each of the four emotions. They reported a performance of F_1 of 0.733 for NMF, which was their highest performing model. One advantage of this approach was that it is unsupervised, which means both that significant amounts of training data are not required and that all the annotated data can be used for testing. This is important because of the small size of the corpus on which the technique was tested.

3 Emotion Recognition Framework

We now describe an unsupervised system for emotion recognition modeled on that reported by Kim

[2] http://www.affective-sciences.org/researchmaterial

Citation	Corpus	Lexicon	# Emotions	Method	F_1
Kim et al. (2010)	Fairy tales	WNA	4	NMF	0.73
Bandhakavi et al. (2017)	Tweets	UMM+DSEL	6	Lexicon only	0.64
Aman and Szpakowicz (2008)	Blog	-	6	Unigrams	0.57
Cherry et al. (2012)	Suicide notes	-	15	SVM+LS	0.55
Strapparava and Mihalcea (2008)	Headlines	-	6	LSA	0.17
Tokuhisa et al. (2008)	"EP" Events	JEE Dict.	10	SVM+kNN	0.5–0.8 Acc.

Table 2: Emotion recognition approaches on narrative-like text, ordered by performance. LSA = Latent Semantic Analysis; LS = Latent sequence modeling

et al. (2010). While we follow the general pattern of that work, we experiment with a different set of dimension reduction methods (NMF from Lee and Seung, as well as PCA and LDA). The system takes as input the following items:

- A corpus containing n sentences $S : s_1, s_2, \ldots, s_n$;
- A set of emotions $E = \{e_1, e_2, \ldots, e_{l-1}, \text{neutral}\}$ for classifying emotions into l different classes, including neutral; and,
- An emotion lexicon $L : \Omega \mapsto E$ which maps each word in the corpus $\omega \in \Omega$ (where Ω has m terms) to an emotion $e \in E$. The word ω is in its lemmatized form and has a specific POS.

A flowchart of the system is shown in Figure 2. The system comprises four consecutive steps. In the first step, **pre-processing**, the system processes the input corpus using the CoreNLP library (Manning et al., 2014) to separate the text into sentences and lemmatized tokens. The second step, **vector space modeling**, uses the lemmatized tokens to generate a vector for each sentence in a vector space whose dimensions correspond to the items in Ω. In the third step, **noise cancellation or dimension reduction**, we explored three different models (Non-negative Matrix Factorization, Latent Dirichlet Allocation, and Principal Component Analysis) to either reduce dimensions or extract features of the vector space. One of our main contributions here is to analyze and explain the effect of this step on the performance of the final emotion recognition system. Finally, the fourth step, **labeling**, compares the vector for each sentence with vectors for each emotion, choosing the closest emotion as the label for the sentence.

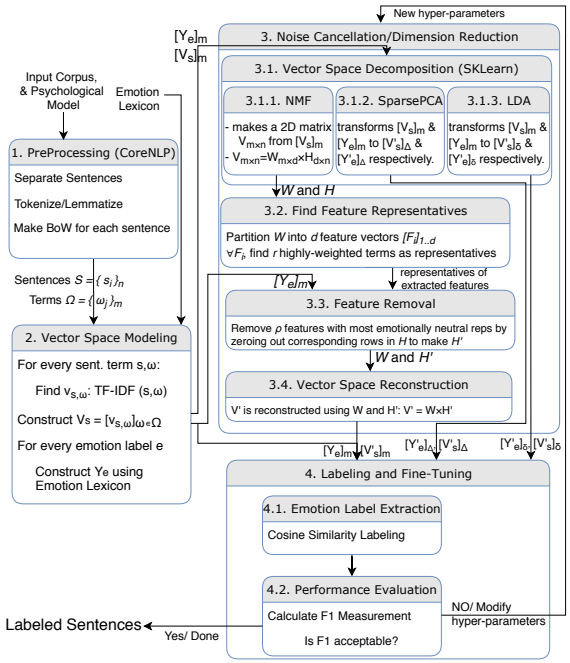


Figure 2: Flowchart of the proposed system. $[V_s]_m$ and $[Y_e]_m$ represent the original m-dimensional sentence and emotion vector model respectively, $[V'_s]_m$, $[V'_s]_\Delta$ and $[V'_s]_\delta$ denote the transformed sentence vector model using NMF, PCA and LDA techniques respectively. $[Y'_e]_\Delta$ and $[Y'_e]_\delta$ denote the transformed emotion vector model using PCA and LDA techniques respectively.

Augmenting WNA As mentioned before, WNA 1.1 assigns an emotion label to 1,471 synonym sets (synsets) of WordNet. This corresponds to a lexicon of nearly 3,495 affective lemma-POS pairs. Careful inspection of WNA revealed both incorrectly included as well as missing pairs. For incorrectly included pairs, a substantial number were included because all their multiple senses were labeled by emotions related to a secondary affective sense, not their main non-affective sense. We manually reviewed and removed these incorrect labels. Additionally, we identified missing lemma-POS pairs with the help of closely related pairs already labeled by WNA. For example the pair *glorious-JJ* was missing from WNA, but is related (via the *derived-from* relation) to already labeled pair *glorify-VB*. We manually searched for these missing relationships, adding the missing terms, as well as recursively adding their synonyms (e.g.,

glorious-JJ resulted in *splendid*, *magnificent*, *brilliant*, and *superb* being added as well). In total, we removed 613 and added 814 labels of different lemma-POS pairs, resulting a final count of 4048 lemma-POS pairs.

In general, the technique of using a fixed lexicon of emotion terms to capture highly context-dependent emotional expressions is problematic at best. Although we show here that work on improving the lexicon does improve emotion recognition results, ultimately, any technique will have to move away from a rigid lexicon-based approach to something more flexible. We plan to explore such directions in future work.

Step 1: Pre-Processing

For each sentence $s \in S$ in the given corpus, we construct a bag of words by tokenizing the sentence and lemmatizing each word. We generate a count vector for BoW_s by mapping each lemma to the count in the sentence ($\Omega \mapsto \mathbb{Z}_{\geq 0}$). We do not remove stop words as their effects are minimized by the *tf-idf* computation in the next step.

Step 2: Vector Space Modeling

Using the count vectors constructed in the first step, we compute a *tf-idf* vector for each sentence as well as a standard vector for each emotion class $e \in E$. For each sentence $s_j \in S$, we construct an m dimensional vector where each entry in the vector is the *tf-idf* of term ω_i in sentence s_j; i.e.

$$v_{ij} = \text{TF}_{i,j} \times \text{IDF}_i \quad (1)$$

where $\text{TF}_{i,j} = \text{BoW}_{s_j}(\omega_i)$,

$$\text{IDF}_i = \log \frac{n}{|\{s \in S : BoW_s(\omega_i) > 0\}|}. \quad (2)$$

n is the number of sentences, and $\Omega = \{\omega_i\}_{i=1}^{m}$.

The constructed vector space model is represented by the following $m \times n$ matrix V:

$$V = [V_{s_1} V_{s_2} \ldots V_{s_n}] \text{ where } V_{s_j} = \begin{pmatrix} v_{1j} \\ v_{2j} \\ \vdots \\ v_{mj} \end{pmatrix} \quad (3)$$

We compute a standard vector for each emotion class $Y_e = (y_{e,\omega_1}, y_{e,\omega_2}, \ldots, y_{e,\omega_m})$ where y_{e,ω_i} is 1 if the term ω_i is mapped to e by the lexicon, otherwise 0.

Step 3: Noise Cancellation or Dimension Reduction

The vectors V_s and Y_e from the previous step are all m-dimensional vectors where m is the total number of terms in the corpus. There are many terms that have little or no effect on the emotion labeling of their sentences. Therefore, dimensional reduction or noise cancellation techniques may improve the performance of the emotion labeling step which comes later. Principle Component Analysis (PCA) has been known for quite some time for noise cancellation (Abdi and Williams, 2010), while Latent Dirichlet Allocation (LDA) was specifically developed for dimension reduction in natural language processing (Blei et al., 2003). Non-Negative Matrix Factorization (NMF) was first introduced for noise cancellation by Lee and Seung (1999).

Step 3.1: Vector Space Decomposition

We can decompose the obtained matrix V in one of the following three ways:

1. Non-negative Matrix Factorization (NMF): we extract d features from the m-dimensional vectors of sentences using NMF.

2. Principal Component Analysis (PCA): We reduce the number of dimensions of V_s vectors from m to $\Delta < m$.

3. Latent Dirichlet Allocation (LDA): We reduce the number of dimensions of V_s vectors from m to $\delta < m$.

When using PCA or LDA we can move directly to fourth step of the system; however, in the case of NMF, we must select important terms (Step 3.2), remove irrelevant features (Step 3.3), and reconstruct the vector space (Step 3.4).

When using NMF for decomposing the vector space model, V is factorized into two matrices $W_{m \times d} = [w_{ij}]$ and $H_{d \times n} = [h_{ij}]$, both with all non-negative entries:

$$V = W \times H \text{ s.t. } w_{ij} \geq 0 \text{ and } h_{ij} \geq 0 \quad (4)$$

Note that d is considered a hyper-parameter in this step and its numerical value can be fine-tuned by maximizing the output of the system on a development set.

The NMF factorization process produces a matrix W whose d columns each represents an m-dimensional feature for each of the original n sentences in the corpus:

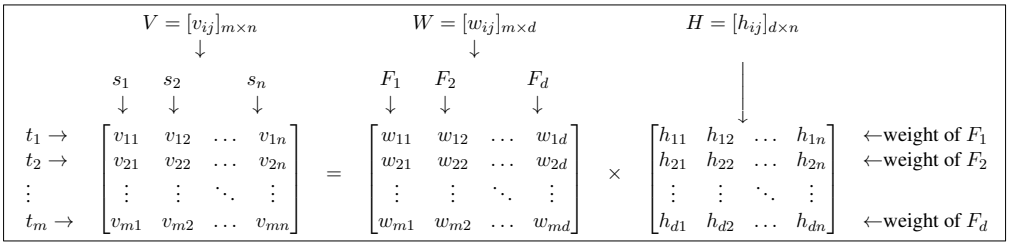


Figure 3: Non-negative matrix factorization (Step 3.1) to extract features of sentence vector model V. The results of this process is given by matrices W and H. Columns of W are corresponding to the extracted features $F_1, F_2, \dots, F_d$ of the model and rows of H are called the weights of these features.

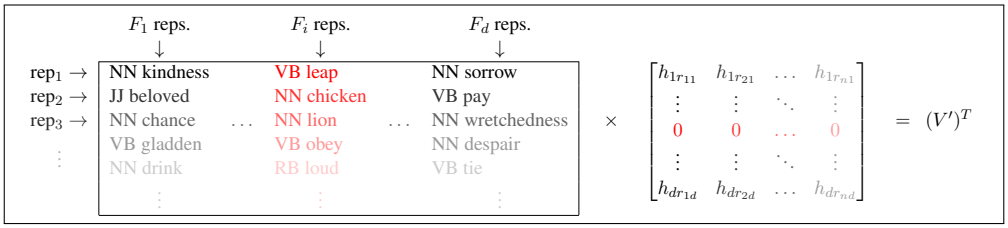


Figure 4: The least relevant features are removed by zeroing out their corresponding weights in matrix H. The updated H matrix is denoted by H'. The sentence vector model is then reconstructed by multiplying W by H' (Steps 3.3 & 3.4). The updated sentence vector model is represented by matrix V'.

$$W = [F_1 F_2 F_3 \dots F_d] \text{ where } F_j = \begin{pmatrix} w_{1j} \\ w_{2j} \\ \vdots \\ w_{mj} \end{pmatrix} \quad (5)$$

Each of the d rows of H matrix represents weights of the d features in F. This decomposition is shown in Figure 3.

Step 3.2: Term Selection

For every feature F_j, we identify a fraction r of terms with the highest weights as its representatives, where r is a hyper-parameter that can be fine-tuned during system optimization (r is usually less than 1%).

Step 3.3: Feature Removal

In this phase we remove the ρ features that have little or no emotional relevance, where ρ is a non-negative integer hyper-parameter that can be tuned. We will call a feature "emotionally irrelevant" if all of its representative terms (as selected in the previous step) are labeled as neutral by the lexicon. These features will always be removed first. If ρ is less than the number of emotionally irrelevant features, we choose at random. On the other hand, if the number of emotionally irrelevant features is less than ρ, we eliminate features F_j in order of their overall emotional relevance, which is computed by estimating the standard deviation of cosine similarity ratios between emotion vectors Y_e's obtained in Step 2 and $F_j \circ R_j$ (element-wise product of F_j and R_j) where R_j is the binary identifier of whether a term is a representative for F_j and is constructed based on the outcome of Step 3.2. Symbolically, to quantify how emotionally relevant feature F_j is, we calculate the following standard-deviation:

$$\sigma_j = \text{StdDev}_{e\in E\setminus\text{neutral}}\{\text{sim}_{\cos}(Y_e, F_j \circ R_j)\} \quad (6)$$

Step 3.4: Vector Space Reconstruction

In this step, the vector space model is reconstructed (V') after eliminating the irrelevant features. Let I denote the set of indices whose corresponding features are identified as least relevant in previous step. Then the reconstructed vector space is:

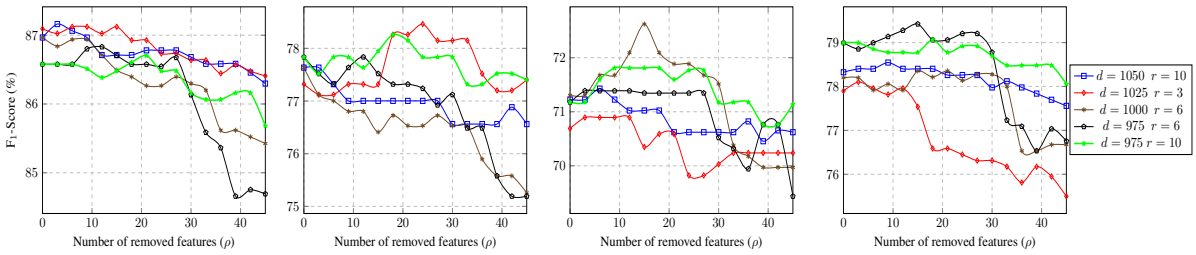
$$V' = [v'_{ij}]_{m\times n} \text{ s.t. } v'_{ij} = \sum_{\substack{1\le k\le d \\ k\notin I}} w_{ik} h_{kj} \quad (7)$$

Figure 4 illustrates the vector space reconstruction.

Step 4: Labeling

Finally the emotion recognition process takes place by measuring the similarity between sentence vectors V_s and standard emotion vectors Y_e which are taken from the previous step with the help of NMF, PCA, or LDA. Label of each sentence s is calculated by the following formula:

$$\text{predicted label of } s = \arg\max_{e\in E} \text{ sim}(V_s, Y_e) \quad (8)$$



(a) Joy (b) Anger (c) Fear (d) Sadness

Figure 5: Exploration of the hyper-parameter space for NMF. Each combination of hyper-parameters d, r, and ρ (dimensions, representatives, and removed features) results in a specific F_1 score for each emotion label. The model with $(d, r) = (975, 10)$, highlighted with green color, results in the highest overall F_1 score when $\rho = 18$. For each individual emotion, the best F_1 score is found at (a) Joy: $(d, r, \rho) = (1050, 10, 3)$, (b) Anger: $(d, r, \rho) = (1025, 3, 24)$, (c) Fear: $(d, r, \rho) = (1000, 6, 15)$, (d) Sadness: $(d, r, \rho) = (975, 6, 15)$.

where similarity function can be measured by the cosine of angle made by the two given vectors:

$$\text{sim}_{\cos}(V_s, Y_e) = \frac{V_s \cdot Y_e}{||V_s|| \times ||Y_e||} \quad (9)$$

4 Performance on Fairy Tale Data

We tuned and tested our system using the manually annotated dataset of fairy tales constructed by Alm (2008), which comprises 176 children's fairy tales (80 from Brothers Grimm, 77 from Hans Andersen, and 19 from Beatrix Potter) with 15,087 unique sentences (15,302 sentences), 7,522 unique words and 320,521 total words. These fairy tales were annotated by two annotators labeling the emotion and mood of each sentence as one of joy, anger, fear, sadness, or neutral which resulted in four labels per sentence. Across the sentences, only 1,090 of them agreed on *all four non-neutral labels*. Kim et al. (2010) used only these sentence to train and test their system[3], and we followed the same procedure. There were 2,405 unique term-POS pairs. Also, the distribution of labels in the dataset is specified in the pie-chart depicted in Figure 6.

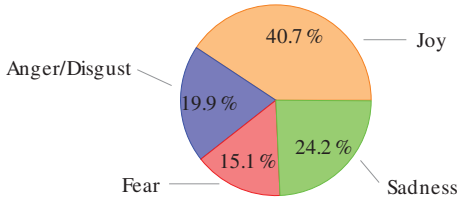


Figure 6: Fairy tales label distribution of sentences with unanimous inter-annotator agreement.

We measured the performance of our system on Alm's data. Without augmenting WNA, using the original 1,471 synsets of WNA, the F_1 score is 0.625. The performance metrics presented in Table 4 were obtained by the model using the augmented WNA. The plots depicted in Figure 5 show the F_1 scores of various setups of the proposed model using NMF technique for noise cancellation. Also, Table 4 summarizes the precision, recall and F_1 score of our system for each of the four emotion classes as well as its overall F_1 score when using NMF, PCA, or LDA with different setups (values of hyper-parameters). As observed in this table, the highest overall F_1 score is obtained when using NMF with $(d, r, \rho) = (975, 10, 18)$. In this model, 209 sentences were labeled incorrectly. Among them, some challenging examples are in Table 3.

5 Unsolved Challenges and Future Work

As already discussed, one challenge regarding automatic emotion recognition is the context dependency of emotional semantics. For instance, *I'm over the moon!* is an expression of extreme happiness but does not use any explicitly happy or joyful words (or, indeed, any emotion word at all). Another obstacle is polysemous words, when words have both an emotional and non-emotional senses; recognizing which sense of the word is being used is challenging and remains an open problem. Aside from these fundamental issues, there is a serious lack of high-quality annotated data, not just for narrative text but for all discourse types. Annotated corpora use a wide variety of sometimes incompatible emotion theories and are often poorly annotated, with low inter-annotator agreements and many errors.

Given these considerations, there are many possible directions for future work, for example:

- Reconciling emotion lexicons and context de-

[3] Kim et al. (2010) reported 1,093 sentences, but we found and removed three sentences that were repeated in the data.

Sentence	Predicted	Gold Label
They told him that their father was very ill, and that they were afraid nothing could save him.	Fear	Sadness
And in sight of the bridge! Said poor pigling, nearly crying.	Sadness	Fear
She smiled once more, and then people said she was dead.	Sadness	Joy
Then he aimed a great blow, and struck the wolf on the head, and killed him on the spot! . . . and when he was dead they cut open his body, and set Tommy free.	Anger	Joy

Table 3: Challenging examples of sentences incorrectly labeled by the model with the most accurate settings.

		Joy			Anger			Fear			Sadness			Overall	
Method	Setup	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	F_1	Acc.
NMF	1050,10,3	0.872	0.872	**0.872**	0.878	0.696	0.776	0.672	0.758	0.712	0.753	0.818	0.784	0.807	0.806
	1025,3,24	0.859	0.876	0.867	0.884	0.705	**0.785**	0.682	0.715	0.698	0.733	0.799	0.764	0.800	0.799
	1000,6,15	0.872	0.858	0.865	0.861	0.687	0.764	0.692	0.764	**0.726**	0.742	0.830	0.784	0.804	0.803
	975,6,15	0.860	0.874	0.867	0.882	0.691	0.775	0.689	0.739	0.713	0.759	0.833	**0.794**	0.808	0.807
	975,10,18	0.858	0.874	0.866	0.879	0.705	0.783	0.703	0.733	0.718	0.755	0.830	0.791	**0.809**	**0.808**
PCA	1050	0.884	0.775	**0.826**	0.760	0.700	0.729	0.552	0.770	0.643	0.756	0.777	0.766	**0.760**	**0.689**
	1150	0.885	0.764	0.820	0.743	0.719	**0.731**	0.542	0.745	0.628	0.748	0.765	0.757	0.752	0.683
	950	0.883	0.766	0.820	0.722	0.696	0.709	0.571	0.782	**0.660**	0.759	0.777	0.768	0.757	0.686
	1100	0.888	0.768	0.824	0.744	0.710	0.726	0.542	0.745	0.628	0.765	0.788	**0.776**	0.758	0.684
LDA	1650	0.636	0.768	**0.696**	0.597	0.498	0.543	0.414	0.424	0.419	0.603	0.466	0.526	0.589	0.589
	1350	0.598	0.791	0.681	0.651	0.558	**0.600**	0.482	0.333	0.394	0.522	0.402	0.454	0.581	0.581
	1300	0.584	0.809	0.678	0.566	0.475	0.516	0.594	0.461	**0.519**	0.570	0.356	0.438	0.580	0.580
	2350	0.671	0.640	0.655	0.524	0.498	0.511	0.456	0.497	0.475	0.584	0.621	**0.602**	0.585	0.585
	1700	0.652	0.696	0.673	0.622	0.516	0.564	0.454	0.533	0.490	0.603	0.553	0.577	**0.601**	**0.601**

Table 4: Comparison of accuracy quantifiers of different models for detecting different emotions. The upper part of the table shows performance of the proposed model using NMF technique with different values of (d, r, ρ); while the middle and bottom parts determine the model accuracy when PCA and LDA techniques are used respectively. The highest F_1 scores of each noise cancellation technique are highlighted.

pendency of emotion detection models using learning techniques;

- Evaluating the performance of a bag-of-words multi-layer perceptron applied to the dataset to extract emotions;
- Applying multi-label prediction to the dataset and comparing the results with this work,
- Evaluating the effect of text unit size (sentence, paragraph, story) on the accuracy of sentiment labels; i.e., would there be an advantage in grouping sentences into longer units (e.g. paragraphs) and assigning a single label to this longer unit? It seems that a sentence by itself might not always carry sufficient cues to disambiguate its emotion, but its surrounding sentences might give this context.

6 Contributions

We identified a high performing approach to emotion recognition in narrative text (Kim et al., 2010) and carefully reimplemented and characterized the technique, exploring a design space of three different noise cancellation or dimension reduction techniques (NMF, PCA, or LDA), exploring various hyper-parameter settings. Our experiments indicated that NMF performed best, with an overall F_1 of 0.809. In the course of our investigation we clarified numerous implementational issues of the work reported by Kim et al. (2010), as well as made some improvements to WordNet Affect (WNA), one of the language resources used in the system, by adding new terms manually and using Wordnet similarity relations. This work suggests several promising future directions for improving the work, including careful annotation of a larger corpus, and augmenting WNA or similar lexicons to provide improved coverage of emotion terms. We release our code and data to enable future work[4].

Acknowledgments

This material is based upon work supported by the U.S. Department of Homeland Security under Grant Award Number, 2017-ST-062-000002. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security.

[4]Code and data can be downloaded from https://doi.org/10.34703/gzx1-9v95/03RERQ

References

Hervé Abdi and Lynne J. Williams. 2010. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2:433–459.

Ebba Cecilia Ovesdotter Alm. 2008. *Affect in *Text and Speech*. Ph.D. thesis, University of Illinois at Urbana-Champaign, Urbana-Champaign, IL.

Saima Aman and Stan Szpakowicz. 2008. Using roget's thesaurus for fine-grained emotion recognition. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*, pages 312–318, Hyderabad, India.

Anil Bandhakavi, Nirmalie Wiratunga, P Deepak, and Stewart Massie. 2014. Generating a word-emotion lexicon from #emotional tweets. In *Proceedings of the Third Joint Conference on Lexical and Computational Semantics (*SEM 2014)*, pages 12–21, Dublin, Ireland.

Anil Bandhakavi, Nirmalie Wiratunga, Deepak Padmanabhan, and Stewart Massie. 2017. Lexicon based feature extraction for emotion text classification. *Pattern Recognition Letters*, 93:133–142.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022.

Christos Boutsidis and Efstratios Gallopoulos. 2008. SVD based initialization: A head start for nonnegative matrix factorization. *Pattern Recognition*, 41(4):1350–1362.

Margaret M Bradley and Peter J Lang. 1999. Affective Norms for English Words (ANEW): Instruction manual and affective ratings. Technical Report C-1, The Center for Research in Psychophysiology, Gainseville, FL.

Fabio Calefato, Filippo Lanubile, and Nicole Novielli. 2017. EmoTxt: a toolkit for emotion recognition from text. In *Proceedings of the Seventh International Conference on Affective Computing and Intelligent: Interaction Workshops and Demos (ACIIW 2017)*, pages 79–80, San Antonio, TX.

Rafael A Calvo and Sunghwan Mac Kim. 2013. Emotions in text: Dimensional and categorical models. *Computational Intelligence*, 29(3):527–543.

Erik Cambria. 2016. Affective computing and sentiment analysis. *IEEE Intelligent Systems*, 31(2):102–107.

Erik Cambria, Andrew Livingstone, and Amir Hussain. 2012. The hourglass of emotions. In Anna Esposito, Antonietta M. Esposito, Alessandro Vinciarelli, Rüdiger Hoffmann, and Vincent Müller, editors, *Cognitive Behavioural Systems*, pages 144–157. Springer, Berlin. Published as Volume 7403, Lecture Notes in Computer Science (LNCS).

Colin Cherry, Saif M Mohammad, and Berry de Bruijn. 2012. Binary classifiers and latent sequence models for emotion detection in suicide notes. *Biomedical Informatics Insights*, 5:BII–S8933.

Roddy Cowie, Ellen Douglas-Cowie, and Cate Cox. 2005. Beyond emotion archetypes: Databases for emotion modelling using neural networks. *Neural Networks*, 18(4):371–388.

Charles Darwin and Phillip Prodger. 1998. *The Expression of the Emotions in Man and Animals*. Oxford University Press, Oxford, UK.

Paul Ekman. 1992a. Are there basic emotions? *Psychological Review*, 99(3):550–553.

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

Paul Ekman. 1993. Facial expression and emotion. *American psychologist*, 48(4):384.

Christiane Fellbaum, editor. 1998. *WordNet: an electronic lexical database*. MIT Press, Cambridge, MA.

Johnny RJ Fontaine, Klaus R Scherer, Etienne B Roesch, and Phoebe C Ellsworth. 2007. The world of emotions is not two-dimensional. *Psychological Science*, 18(12):1050–1057.

Carroll E Izard. 2007. Basic emotions, natural kinds, emotion schemas, and a new paradigm. *Perspectives on Psychological Science*, 2(3):260–280.

William James. 1890. *The Principles of Psychology*. Henry Holt and Company, New York.

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, Los Angeles, CA.

Matthew W Kreuter, Melanie C Green, Joseph N Cappella, Michael D Slater, Meg E Wise, Doug Storey, Eddie M Clark, Daniel J O'Keefe, Deborah O Erwin, Kathleen Holmes, et al. 2007. Narrative communication in cancer prevention and control: a framework to guide research and application. *Annals of Behavioral Medicine*, 33(3):221–235.

George Langroudi, Anna Jourdanous, and Ling Li. 2018. Music emotion capture: Sonifying emotions in eeg data. In *Symposium on Emotion Modeling and Detection in Social Media and Online Interaction*, pages 1–4, Liverpool, UK.

Daniel D Lee and H Sebastian Seung. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791.

Hugo Lövheim. 2012. A new three-dimensional model for emotions and monoamine neurotransmitters. *Medical Hypotheses*, 78(2):341–348.

Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. 2009. Online dictionary learning for sparse coding. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 689–696, Montreal, Canada.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014): System Demonstrations*, pages 55–60, Baltimore, MD, U.S.

DA Medler, A Arnoldussen, JR Binder, and MS Seidenberg. 2005. The Wisconsin Perceptual Attribute Ratings (WPAR) database. Retrieved from http://www.neuro.mcw.edu/ratings on April 23, 2020.

Saif M Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. *arXiv preprint arXiv:1308.6242*.

Saif M Mohammad and Peter D Turney. 2010. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 26–34, Los Angeles, CA.

Keith Oatley and Philip N Johnson-Laird. 1987. Towards a cognitive theory of emotions. *Cognition and Emotion*, 1(1):29–50.

Andrew Ortony, Gerald L Clore, and Allan Collins. 1990. *The Cognitive Structure of Emotions*. Cambridge University Press, Cambridge, UK.

Jaak Panksepp, Brian Knutson, and Douglas L Pruitt. 1998. Toward a neuroscience of emotion. In *What develops in emotional development?*, pages 53–84. Springer, Boston, MA.

W Gerrod Parrott. 2001. *Emotions in Social Psychology: Essential Readings*. Psychology Press, London.

Haoruo Peng, Daniel Khashabi, and Dan Roth. 2019. Solving hard coreference problems. *arXiv preprint arXiv:1907.05524*.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic Inquiry and Word Count (LIWC) Software.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. In Robert Plutchik, editor, *Theories of Emotion*, pages 3–33. Elsevier, Amsterdam, Netherlands.

Robert Plutchik. 1984. Emotions and imagery. *Journal of Mental Imagery*, 8:105–111.

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.

James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

Klaus R Scherer. 2005. What are emotions? and how can they be measured? *Social Science Information*, 44(4):695–729.

Shadi Shaheen, Wassim El-Hajj, Hazem Hajj, and Shady Elbassuoni. 2014. Emotion recognition from text based on automatically generated rules. In *Proceedings of the 2014 IEEE International Conference on Data Mining Workshop*, pages 383–392, Shenzhen, China.

Phillip Shaver, Judith Schwartz, Donald Kirson, and Cary O'connor. 1987. Emotion knowledge: Further exploration of a prototype approach. *Journal of Personality and Social Psychology*, 52(6):1061–1086.

Carlo Strapparava and Rada Mihalcea. 2008. Learning to identify emotions in text. In *Proceedings of the 2008 ACM symposium on Applied computing*, pages 1556–1560, Fortaleza, Ceara, Brazil.

Carlo Strapparava and Alessandro Valitutti. 2004. WordNet Affect: an affective extension of wordnet. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, pages 1083–1086, Lisbon, Portugal.

Jared Suttles and Nancy Ide. 2013. Distant supervision for emotion classification with discrete binary values. In *Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics*, pages 121–136, Berlin, Germany.

Zhi Teng, Fuji Ren, and Shingo Kuroiwa. 2007. Emotion recognition from text based on the rough set theory and the support vector machines. In *Proceedings of the 2007 International Conference on Natural Language Processing and Knowledge Engineering*, pages 36–41, Beijing, China.

Ryoko Tokuhisa, Kentaro Inui, and Yuji Matsumoto. 2008. Emotion classification using massive examples extracted from the web. In *Proceedings of the 22nd International Conference on Computational Linguistics: Volume 1*, pages 881–888, Manchester, UK.

Ozlem Uzuner, Andreea Bodnari, Shuying Shen, Tyler Forbush, John Pestian, and Brett R South. 2012. Evaluating the state of the art in coreference resolution for electronic medical records. *Journal of the American Medical Informatics Association*, 19(5):786–791.

Xiao Zhang, Wenzhong Li, Xu Chen, and Sanglu Lu. 2018. Moodexplorer: Towards compound emotion detection via smartphone sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(4):1–30.