# Lexicon based Feature Extraction for Emotion Text Classification

**Article** *in* Pattern Recognition Letters · December 2016

DOI: 10.1016/j.patrec.2016.12.009

CITATIONS
120

READS
3,527

**4 authors:**

Anil Sriharsha Bandhakavi
Robert Gordon University
**7** PUBLICATIONS **264** CITATIONS
SEE PROFILE

Nirmalie Wiratunga
Robert Gordon University
**141** PUBLICATIONS **1,685** CITATIONS
SEE PROFILE

Deepak P
Queen's University Belfast
**142** PUBLICATIONS **1,039** CITATIONS
SEE PROFILE

Stewart Massie
Robert Gordon University
**66** PUBLICATIONS **875** CITATIONS
SEE PROFILE

Pattern Recognition Letters
journal homepage: www.elsevier.com

# Lexicon based Feature Extraction for Emotion Text Classification

Anil Bandhakavi[a,**], Nirmalie Wiratunga[a], Deepak P.[b], Stewart Massie[a]

[a] *School of Computing Science and Digital Media, Robert Gordon University, Aberdeen and AB10 7QB, UK*
[b] *School of Electronics Electrical Engineering and Computer Science, Queens University, Belfast and BT7 1NN, UK*

ABSTRACT

General Purpose Emotion Lexicons (GPELs) that associate words with emotion categories remain a valuable resource for emotion analysis of text. However the static and formal nature of their vocabularies make them inadequate for extracting effective features for document representation, in domains that are inherently dynamic in nature (e.g. Social Media). This calls for lexicons that are not only adaptive to the lexical variations in a domain but also provide finer-grained quantitative estimates to accurately capture word-emotion associations. In this paper we extend prior work on domain specific emotion lexicon (DSEL) generation and apply it for emotion feature extraction. We demonstrate how our generative unigram mixture model (UMM) based DSEL learnt by harnessing labelled (blogs, news headlines and incident reports) and weakly-labelled (tweets) emotion text can be used to extract effective features for emotion classification. Our results confirm that the features derived using the proposed lexicon outperform those from state-of-the-art lexicons learnt using supervised Latent Dirichlet Allocation (sLDA) and Point-Wise Mutual Information (PMI). Further the proposed lexicon features also outperform state-of-the-art features derived using a combination of n-grams, part-of-speech information and sentiment lexicons.



## 1. Introduction

Emotion is an important factor that influences overall human behaviour which include rational tasks such as reasoning, decision making and interaction. Though emotions are subjective, they occur in objectively deducible ways in text [1]. Emotion analysis concerns the computational study of natural language expressions in order to identify their associations with different emotions such as *anger, fear, joy, sadness, surprise* etc.

Sentiment analysis [2] is widely used to gauge user opinion expressed in text. However it is often desirable to have more detailed information about the views and opinions of people. For example, if users like a brand, insights such as they are positive because the brand gives them joy or they felt surprised by the endorser, offers crucial business intelligence for brand management. Given that there is unprecedented access to emotion-rich content through tweets, blogs and discussion posts there is a great opportunity and need to build automatic tools, in order to understand the emotions of the users. Emotion classification is among the most widely studied problems in emotion analysis of text, where supervised machine learning methods are leveraged to classify text documents [3, 4] into emotion classes, induced from emotion theories proposed in psychology by Ekman [5], Parrot [6] and Plutchik [7]. Among the two approaches for emotion modelling [4], one on discrete emotions and another using the continuum approach, this work builds upon the former. Further the discrete model for emotion detection has been subject to extensive exploration in psychological research [8, 9].

Representation of text documents is a crucial step in machine learning approaches for text classification. A popular representation involves refining the Bag-of-words (BoW) or n-grams feature vector, so that a subset of words are chosen using a selection metric to represent a text document [10]; this is normally referred to as feature selection. Feature engineering, on the other hand, is about building a set of new features rather than selecting a subset of words. Such features could be frequency of higher-level concepts such as *topics* [11], or may use semantic representations derived from an ontology [12]. More specialized tasks call for more fine-tuned feature repre-

**Corresponding author: Tel.: +0-000-000-0000; fax: +0-000-000-0000;
*e-mail:* a.s.bandhakavi@rgu.ac.uk (Anil Bandhakavi)

sentations; for example, the length of contiguous upper-case character sequences is found to be a useful feature for spam filtering[1] whereas the number of lower-cased words resulted in performance improvements in SMS filtering [13]. Author identification is another area where fine-tuned features such as stylemarkers and short-words (e.g. if, is etc.) have enhanced classification accuracy [14]. Emotion Analysis, being as much or even more specialized than the above tasks, has also relied on fine-tuned features described later.

Emotion analysis of text requires careful modelling of text, since words associate with different emotions in different contexts with varying levels of magnitude making the identification of words for document representation more challenging. For example, in a sentence such as *beautiful morning #amazing* the word *beautiful* could be associated moderately with emotions such as *joy* and *love*, *amazing* could be associated strongly with emotion *joy* and *morning* could be weakly associated with emotion *joy*. Such word-emotion associations are usually captured by emotion lexicons. Existing general purpose emotion lexicons (GPELs) such as WordNet-Affect (WNA) [15], EmoSenticNet (ESN) [16] and NRC word-emotion lexicon [17], which are hand crafted, associate between words and emotions identified by Ekman and Plutchik. Emotion features extracted using the knowledge of the GPELs, when combined with traditional BoW features improved emotion classification significantly [18, 19].

However GPELs poorly model the context in which words convey emotions. For example *Glee* might normally connote *joy*, but would need to be assumed neutral in the context of a document corpus talking about the television series with the same name. Further, *unfair* may be associated with *anger* despite being more dominant in *sadness* related documents; the crisp binary memberships of words in GPELs do not allow to capture such fuzzy memberships of words to emotion classes, thereby making them limitedly effective for feature extraction. Accordingly, recent efforts in emotion analysis focused on learning domain specific lexicons [20, 21] and also utilizing them for emotion feature extraction [17, 22]. However the emotion features extracted were limited to simple emotion word counts in a document using the lexicon, which, while being simple, do not exploit the knowledge of the lexicon in its entirety. Accordingly our contributions in this paper are as follows:

- We extend our prior work on domain specific emotion lexicon (DSEL) generation for feature extraction. This is different from our work in [23] since it extracts features beyond simple word counts using the knowledge of a DSEL;
- We introduce novel feature extraction methods to harness the emotion rich knowledge being captured by our DSEL. Unlike our work in [24], which uses DSEL as a direct tool for word and phrase-level emotion analysis, here we extract features to classify text into emotion classes using machine learning; and
- Evaluate through a comparative study the effectiveness of the proposed emotion features on benchmark emotion classification datasets.

In the rest of the paper we review the most related literature in Section 2. In Section 3 we formulate the methods for extracting emotion lexicon based features. Section 4 describes the state-of-the-art baseline features used in our comparative study. In section 5 we describe our experimental set up and analyse the results. Section 6 presents our conclusions and future directions.

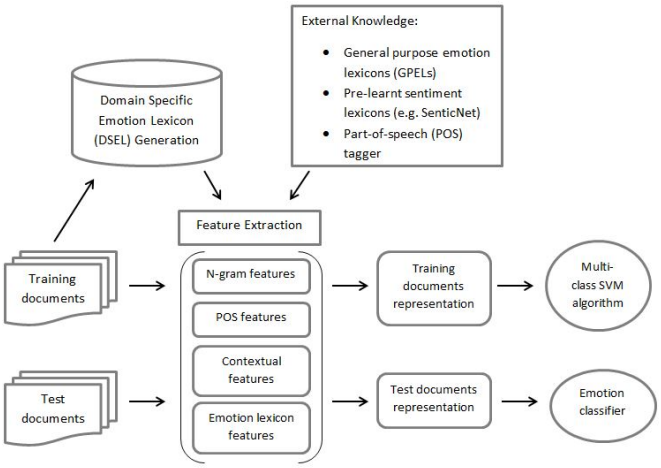


Fig. 1: Feature extraction and emotion classifier learning. A DSEL is learnt from the training documents. Feature extraction is done using standard methods, knowledge of the learnt DSEL and other external resources. An SVM emotion classifier learnt from the training documents is used to classify the test documents.

## 2. Related Work

A majority of the literature concerning emotion analysis is shaped by machine learning approaches. These approaches represent documents as vectors in a feature space and classify them into predefined emotion categories defined by emotion theories such as Ekman [5], Plutchik [7]. The feature extraction process for emotion classification is summarized in figure 1. Observe that the lexicon based features proposed in this work are extracted using the knowledge of the DSEL learnt on the training documents. POS taggers, sentiment lexicons and GPELs act as external resources for extracting relevant features for emotion classification. In the rest of the section we review the state-of-the-art features proposed for emotion classification of text and also relevant literature on emotion lexicon generation.

### 2.1. Features for Emotion Classification

*Generic n-gram features:* This is the most standard representation used in text classification tasks including emotion classification. Documents are represented in a space of unordered list

[1] http://archive.ics.uci.edu/ml/datasets/Spambase

of terms (BoW or n-grams)as vectors. Burget et al. [25] used n-grams with tf-idf weighting [26] to classify Czech news headlines. Similar to the findings in sentiment classification [27], Aman et al. [28] and Matthew et al. [29] demonstrated the effectiveness of n-gram features with binary weighting (word presence/absence) in emotion classification of blogs and tweets respectively. However a common limitation of n-gram features is their inability to capture the underlying emotion semantics, thereby resulting in overall performance degradation. This has lead to research [30, 31] which explores richer features that are better suited for emotion classification.

*Special n-gram features*: As alluded to earlier, specialized features (e.g. punctuation) have been explored in the case of emotion analysis, as in the case of other specialized tasks such as author identification. These features were designed to capture the emotive expressions that occur in subtle ways, especially in Twitter. For instance, Wang et al. [32] designed features such as positional n-grams (i.e. n-grams in the first half of a tweet and n-grams in the second half of a tweet) and part-of-speech (POS) tagging to complement generic n-grams for emotion classification of tweets. Similar to the findings in sentiment classification [27] positional n-grams decreased performance, whilst POS information lead to marginal improvements over n-grams in emotion classification. Roberts et al. [33] found that modelling the presence/absence of punctuation (!, ?) marginally improves classification performance for emotions such as *surprise* and *joy* on tweets.

*Lexicon based features*: These features were designed based on the intuition that sentiment/emotion bearing words identified by lexicons can form useful knowledge to represent documents for emotion classification. Aman et al. [28] augmented generic n-grams with features to count the occurrences of emotion words provided by GPELs to significantly improve emotion classification of blogs. Whilst GPELs offer useful knowledge about emotion-rich words, they are static and are likely to have poor coverage of the emotion vocabulary used in domains like Twitter. For emotion classification of tweets, Mohammad [17] and [22] demonstrated that DSEL based features offer significant gains over n-grams when compared to those of GPEL based features [32]. However feature extraction using DSELs has not been explored beyond binary and integer counts. In particular the knowledge of a DSEL to quantify the association between words and emotions can be leveraged to design more sophisticated features for emotion classification, which is the focus of this work.

*Additional features*: Apart from the aforementioned features additional knowledge sources such as emotion hashtags [34], emotion word lists [30], topic scores [33] were used to design features that complement the n-gram features and general purpose lexicons such as WordNet-Affect. Performance improvements were observed in emotion classification tasks over using n-grams alone [34], but were found to be less effective when compared with lexicon based features suggesting that lexicon based features need to be explored further to design better and more effective text representations for emotion classification of text. In this work we further explore the potential of DSELs to extract effective representations for emotion classification. We also evaluate the contributions of the proposed features by comparing their performance with existing state-of-the-art features in emotion classification.

Table 1: Sample terms in WordNet-Affect Lexicon. The first column in each row denotes an Ekman emotion and the second column in each row denote the words that connote the emotion.

| Emotion | Words |
|---|---|
| Anger | irascibility, short-temper, spleen |
| Fear | frighten, fright, scare |
| Joy | hilarious, screaming, uproarious |
| Sadness | penitently, penitentially |
| Surprise | astonishing, astounding, staggering |
| Disgust | detestably, repulsively, abominably |

### 2.2. Learning emotion lexicons

Existing methods for learning DSELs are mostly supervised, since they rely either on labelled or weakly-labelled emotive content in a domain. Different learning methodologies such as Point-wise Mutual Information (PMI) [22], Latent Dirichlet Allocation (LDA) in a semi-supervised setting [35] have been applied to learn DSELs. In addition, supervised LDA (sLDA) [36] offers a more accurate means to model emotion classes as topics and for lexicon generation. Further crowd-annotated emotional news articles[2] were leveraged for lexicon generation, by combining the document-frequency distributions of words and the emotion distributions over documents [21, 37]. In our prior work [23, 24], we jointly modelled the emotionality and neutrality of words using a unigram mixture model (UMM) to learn DSELs from labelled and weakly-labelled emotion documents.

## 3. Lexicon based Feature Extraction

In this section we first characterize the knowledge offered by GPELs and the DSELs, followed by a brief illustration of our proposed DSEL generation process. Thereafter we formulate the different emotion relevant features that can be extracted using the knowledge of GPELs and DSELs.

### 3.1. Emotion Lexicon Knowledge

A GPEL, $Lex(w, j)$ is a list of words per emotion class:

$$Lex(w, j) = \begin{cases} 1 & \text{if } w \in List(e_j), \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where $List(e_j)$ denotes the list of words corresponding to the $j^{th}$ emotion in the GPEL. A sample of the GPEL, WNA is shown in table 1

In contrast to GPELs, a DSEL quantifies the associations between words in a vocabulary $V$ and a set of pre-defined emotions $E$. For any given arbitrary word $w$, the dominant emotion $e$ expressed is calculated using the lexicon as follows:

$$e = \arg\max_j Lex(w, e_j) \quad (2)$$

[2] http://www.rappler.com/

A DSEL in our case is learnt from a corpus of emotion labelled documents by jointly modelling emotionality and neutrality of words using a generative unigram mixture model (UMM). In the following section we briefly explain our proposed lexicon generation method. Further details about our proposed DSEL generation can be found in [23, 24].

### 3.1.1. Mixture Model for Lexicon Generation

We model real-world emotion data to be a mixture of emotion bearing words and emotion-neutral (background) words. More formally our generative model is as follows to describe the generation of documents connoting emotion $e_t$:

$$P(D_{e_t}, Z|\theta_{e_t}) = \prod_{i=1}^{|D_{e_t}|} \prod_{w \in d_i} [(1 - Z_w)\lambda_{e_t} P(w|\theta_{e_t}) + (Z_w)(1 - \lambda_{e_t})P(w|N)]^{c(w,d_i)} \quad (3)$$

where $\theta_{e_t}$ is the emotion language model and $N$ is the background language model. $\lambda_{e_t}$ is the mixture parameter and $Z_w$ is a binary hidden variable which indicates the language model ($\theta_{e_t}$ or $N$) that generated the word $w$. Further $c(w, d_i)$ is the number of times word $w$ occurs in document $d_i$.

The estimation of parameters $\theta_{e_t}$ and $Z$ can be done using expectation maximization (EM), which iteratively maximizes the complete data ($D_{e_t}$, $Z$) by alternating between E-step and M-step. The E and M steps in our case are as follows:

**E-step**:

$$P(Z_w = 0|D_{e_t}, \theta_{e_t}^{(n)}) = \frac{\lambda_{e_t} P(w|\theta_{e_t}^{(n)})}{\lambda_{e_t} P(w|\theta_{e_t}^{(n)}) + (1 - \lambda_{e_t})P(w|N)} \quad (4)$$

**M-step**:

$$P(w|\theta_{\theta_{e_t}}^{(n+1)}) = \frac{\sum_{i=1}^{|D_{e_t}|} P(Z_w = 0|D_{e_t}, \theta_{e_t}^{(n)})c(w, d_i)}{\sum_{w \in V} \sum_{i=1}^{|D_{e_t}|} P(Z_w = 0|D_{e_t}, \theta_{e_t}^{(n)})c(w, d_i)} \quad (5)$$

where $n$ indicates the EM iteration number. EM is used to estimate the parameters of the $k$ mixture models corresponding to the emotions in $E$. The emotion lexicon $UMMlex$ is learnt by using the $k$ emotion language models and the background model $N$ as follows:

$$UMMlex(w_i, \theta_{e_j}) = \frac{P(w_i|\theta_{e_j}^{(n)})}{\sum_{t=1}^{k} [P(w_i|\theta_{e_t}^{(n)})] + P(w_i|N)} \quad (6)$$

$$UMMlex(w_i, N) = \frac{P(w_i|N)}{\sum_{t=1}^{k} [P(w_i|\theta_{e_t}^{(n)})] + P(w_i|N)} \quad (7)$$

where $k$ is the number of emotions in the corpus, and $UMMlex$ is a $|V| \times (k+1)$ matrix. A sample of our UMM lexicon is shown in table 2. Observe that non-standard and creative expressions such as :), *good*!! are widely used to convey emotions on social media. Such expressions often intensify the emotionality of the text. Modelling such expressions is critical for social media emotion analysis. Therefore in the text preprocessing phase emoticons (e.g. :) ) and concatenated expressions (e.g. good!!) are tokenized as single-words to capture their association with different emotions.

Table 2: A sample of the UMM word-emotion lexicon. Each row represents a word and its association strength with the different emotions captured as a numerical vector.

| Words | Anger | Fear | Joy | Sadness | Surprise | Neutral |
|---|---|---|---|---|---|---|
| :) | 0.056 | 0.085 | 0.533 | 0.062 | 0.072 | 0.192 |
| good!! | 0.074 | 0.109 | 0.305 | 0.236 | 0.093 | 0.183 |
| #arrogant | 0.332 | 0.173 | 0.057 | 0.131 | 0.150 | 0.157 |

### 3.2. Emotion Lexicon Features

As mentioned in the related work section, previous research suggests that lexicon based features improve emotion classification. In this section we explore how the knowledge of a DSEL can be utilized to extract a range of features relevant for emotion classification. Observe that all the lexicon based feature vectors proposed in this work are of length $|E|$, where $|E|$ is the number of emotion classes in a data set. We consider the following features to represent documents:

1. Total Emotion Count (TEC) [22]: This feature captures the number of words in a document that associate with an emotion. Given a document $d$, its corresponding feature vector is denoted by $d_{TEC}$. The feature value for the $j^{th}$ emotion is computed as follows:

$$d_{TEC}[e_j] = \sum_{w \in d} I(e_j = \arg\max_k Lex(w, k)) \times count(w, d) \quad (8)$$

   $I(.)$ is an indicator function and is set to 1 or 0 when the argument is true or false respectively. $count(w, d)$ is the number of occurrences of word $w$ in document $d$. Note that $TEC$ only captures the popular emotion context of a word suggested by the lexicon (i.e., emotion with highest score in the lexicon). However not all words associate with just a single emotion. For example, even if the word *beautiful* may be associated moderately with both the emotions *joy* and *love*, the $TEC$ emotion feature would force the word to contribute a count of 1 towards either of these emotions (depending on the scores from the lexicon $Lex$) and 0 towards the other. Therefore it is important to develop features that incorporate the relations between a word and multiple emotions.
2. Total Emotion Intensity (TEI): This is the sum of the emotion intensity scores of words present in a document. Unlike the coarse integer counts in $TEC$ features, here word-level emotion intensity scores offered by a DSEL are used to capture the emotional orientation of documents along multiple emotion concepts (classes). Accordingly $d_{TEI}$ is the feature vector corresponding to a document $d$. The feature value for the $j^{th}$ emotion is computed as follows:

$$d_{TEI}[e_j] = \sum_{w \in d} Lex(w, e_j) \times count(w, d) \quad (9)$$

3. Max Emotion Intensity (MEI): Research in Sentiment analysis suggest that high sentiment-bearing terms is indicative of sentiment class of the document regardless of the average score for the document [38]. We expect it to be true for emotion analysis as well. Therefore we consider

the intensity score of the highest emotion-bearing word in the given document. Given a document $d$, and its corresponding feature vector $d_{MEI}$, the feature value for the $j^{th}$ emotion is computed as follows:

$$d_{MEI}[e_j] = \arg\max_{w \in d} Lex(w, j) \tag{10}$$

4. Graded Emotion Count (GEC): We extend the idea of utilizing high intensity emotion words to extract document representations by developing variants of $TEC$ and $TEI$. Both $TEC$ and $TEI$ consider all the words in a document regardless of the intensity with which they convey an emotion. However it is useful to understand the impact of high intensity words on emotion classification. $GEC$ is similar in principle to $TEC$, except that it only captures the number of words in a document that associate with an emotion and over a threshold value $\delta$. Since our proposed DSEL quantifies the association between each word and the set of emotions in the form of a probability distribution, the intensity scores always lie in the interval [0, 1]. We divided this interval into 4 quartiles [0, 0.25), [0.25, 0.5), [0.5, 0.75) and [0.75, 1] respectively. Further we used the three values 0.25, 0.5 and 0.75 as threshold $\delta$ in our experiments. The $GEC$ features extracted using the DSELs are for the above three thresholds. Given a document $d$, and its corresponding feature vector $d_{TEC}$, the feature value for the $j^{th}$ emotion is computed as follows:

$$d_{GEC}[e_j] = \sum_{\substack{w \in d \\ Lex(w,j) \geq \delta}} I(e_j = \arg\max_k Lex(w, k)) \times count(w, d) \tag{11}$$

5. Graded Emotion Intensity (GEI): Similar to $GEC$, we develop a variant of $TEI$, $GEI$ which is the sum of intensity scores of words in a document and over a threshold $\delta$. The thresholds mentioned earlier are used for extracting $GEI$ features using DSELs. Given a document $d$, and its corresponding feature vector $d_{GEI}$, the feature value for the $j^{th}$ emotion is computed as follows:

$$d_{GEI}[e_j] = \sum_{\substack{w \in d \\ Lex(w,j) \geq \delta}} Lex(w, e_j) \times count(w, d) \tag{12}$$

## 4. Baseline Emotion Features

In this section we detail the commonly used features to improve emotion classification. Unlike the features discussed in the previous section these features do not rely on the knowledge of an emotion lexicon. We consider the following:

1. n-grams (n=1): These are the most standard corpus level features used in different classification tasks including sentiment [39] and emotion classification [28]. We used a binary weighting (presence/absence) to construct the feature vector, since it is found to be effective by earlier research in sentiment [27] and emotion classification [29].
2. Part-of-Speech (POS) features: Similar to [32], we used features to model the occurrence of verbs, adverbs, nouns and adjectives in a document. Part-of-speech tagging on non-social media data sets is done using the stanford POS tagger[3], whilst Twitter NLP tool [40] from Carnegie Mellon University was used for tagging social media data sets .
3. Contextual features (CF): Though standard words can convey the emotional intention of the author, additional expressions such as punctuation marks, emoticons are often used on social media to express emotions. Further sentiment bearing words could indicate the emotion in the text and also alter its orientation from positive-emotion(e.g. *joy*) to negative-emotion (e.g. *sadness*) or vice versa. We consider the following contextual features used in sentiment [39] and emotion [32] classification for our comparative study:
   - Capitalized words: This feature counts the number of words in a document with all upper case characters [39].
   - Elongated words: This feature counts the number of words with character repeated two, three or four times [39]. For example *haaappy*.
   - Punctuation: Emotions are intensified on social media using exclamation marks and question marks. Similar to [39], two features were included to model the occurrence of question marks and exclamation marks in a document.
   - Emoticons: Emoticons are facial expressions captured pictorially, and are often used on social media to convey emotions. A binary feature is designed to model the presence/absence of emoticons in a document. The emoticon list is adopted from an earlier work in emotion classification [41].
   - Negation: Though the role of negation is not extensively studied for emotion classification, following its usefulness for sentiment classification [27], we include a feature to model the occurrence of negators in documents. We used a standard list of negators proposed by a popular work in sentiment analysis [38].
   - Sentiment features: Though sentiment and emotion are different by definition [42], prior research in emotion classification [32] explored the role of sentiment knolwedge offered by lexicons. Similary we define two integer valued features, to capture for the number of positive words and negative words observed in a document. However in addition to the sentiment lexicons used in [32] we consider more recent lexicons like SentiwordNet [43], SenticNet [44], NRC HashTag sentiment lexicon[4] and Sentiment140 lexicon[4]. An exhaustive list of positive and negative words is created by merging the aforementioned lexicons to extract sentiment features from the documents.

[3] http://nlp.stanford.edu/software/tagger.shtml
[4] http://saifmohammad.com/WebPages/lexicons.html

## 5. Evaluation

The objective of our evaluation is to comparatively evaluate the quality of different document representations proposed in literature for emotion classification, emotion lexicon based representations extracted using the knowledge of baseline lexicons such as PMI, LDA and the proposed UMM based lexicon. We evaluate the different representations for individual class (emotion) performance and also overall performance. Based on this evaluation we construct hybrid representations by combining the best performing baseline features and the best performing lexicon based features. We combined the best performing features (i.e. baseline, lexicon based features) to construct hybrid features expecting further performance improvements. Our evaluation is carried out through emotion classification tasks on benchmark data sets. Emotion classification is chosen, since it is possible to assess the quality of different models that can classify documents into discrete emotion classes. Significance is reported using a paired one-tailed t-test using 95% confidence (i.e. with p value <= 0.05). Observe that results highlighted in bold indicates best performance obtained on that particular data set.

### 5.1. Datasets

We use four benchmark datasets in our evaluation (see Table 4a). Note that on the training documents of each data set, DSELs are learnt (see figure 1) using WED [21], PMI [22], sLDA [36] and by the proposed method (refer section 3.1) in order to extract the lexicon based emotion features mentioned in section 3.2.

#### 5.1.1. News data set (SemEval-2007)

Consists of 1250 emotional news headlines harnessed for evaluating the connection between emotions and lexical semantics at the SemEval-07 workshop [45]. Each headline was provided with emotion ratings in the range [-100, 100] for the Ekman basic emotions. We used this data set for emotion classification, by considering the highest rated emotion for each headline as the class label. Table 3a shows the distribution of different emotion classes in the training and test sets. The dataset is comparatively small with a considerable skewed class distribution. We are particularly interested to explore how the generative DSEL based features compare to baseline features. We expect that the smaller dataset size combined with the skewed distribution makes this an interesting dataset for comparison purposes.

#### 5.1.2. Twitter Dataset

A collection of 0.28 million emotional tweets[5] crawled from the Twitter search API using tweet identification numbers provided by [32]. The emotion labels in the data set correspond to Parrot's primary emotions [6]. We used this data set for emotion classification (stratified 10-fold cross validation). Table 3b shows the average distribution of the different emotion classes over the 10 folds. As is evident from the table, not all emotions are strongly expressed in this data set. Emotions such as *joy, sadness* are more prominent compared to others like *fear, surprise*. Therefore it would be interesting to see how the different document representations fare in performance amidst such class imbalance.

Table 3: Emotion Datasets

(a) News (SemEval-07)

| Emotion | #TrainSet | #TestSet |
|---|---|---|
| Disgust | 35 | 20 |
| Anger | 67 | 23 |
| Fear | 155 | 33 |
| Surprise | 184 | 38 |
| Sadness | 201 | 61 |
| Joy | 358 | 75 |
| Total | 1000 | 250 |

(b) Twitter

| Emotion | #TrainSet | #TestSet |
|---|---|---|
| Surprise | 2233 | 282 |
| Fear | 12592 | 1548 |
| Love | 30117 | 3464 |
| Anger | 57310 | 6496 |
| Sadness | 62611 | 7069 |
| Joy | 73098 | 8235 |
| Total | 237961 | 27095 |

(c) Blogs

| Emotion | #TrainSet | #TestSet |
|---|---|---|
| Surprise | 91 | 16 |
| Fear | 91 | 41 |
| Sadness | 136 | 57 |
| Anger | 140 | 36 |
| Joy | 416 | 69 |
| Total | 874 | 219 |

(d) Incident reports (ISEAR)

| Emotion | #TrainSet | #TestSet |
|---|---|---|
| Disgust | 815 | 203 |
| Anger | 816 | 204 |
| Fear | 815 | 204 |
| Guilt | 815 | 204 |
| Joy | 815 | 204 |
| Sadness | 815 | 204 |
| Shame | 816 | 203 |
| Total | 5707 | 1426 |

#### 5.1.3. Blog Dataset

Consists of 5500 blog sentences annotated with Ekman basic emotions by 3 annotators with an average inter annotator agreement (kappa of 0.76) [18]. Table 3c shows the average distribution of different emotion classes over the 5 folds. The emotion class distribution is highly skewed towards the emotion *joy*. Further the smaller size of the data set is likely to challenge the modelling of the weakly represented emotions like *fear, surprise*.

#### 5.1.4. Incident reports data set (ISEAR)

Consists of 7000 incident reports obtained from an international survey on emotion reactions [6]. Each report is an emotion summary, describing the situation which lead the participant to experience one of 7 emotions: *anger, disgust, fear, shame, guilt, joy and sadness*. Table 3d shows the average distribution of different emotion classes over the 5 folds. Unlike the other data sets the emotion classes here have a near uniform distribution, which is very unlikely in a real word sample of emotion rich content. It will also be interesting to observe how closely related emotions such as *shame* and *guilt* might be differentiated in the classification task.

### 5.2. Baselines and Metrics

The following document representations are used in our comparative study:

- Baseline emotion features (see section 4);

[5] http://knoesis.org/?q=projects/emotion

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

Table 4: Emotion examples and Lexicon vocabulary statistics

(a) Exdata

| Data set | Example | Emotion |
|---|---|---|
| Twitter | going to las vegas tomorrow :) #excited #happy | Joy |
| Blogs | I'm scared of my future | Fear |
| News (SemEval-07) | Trolley Square shooting leaves 6 dead | Anger |
| Incident reports (ISEAR) | My friend ran away from home,I feel for him | Sadness |

(b) Lexstat

| Lexicon | #terms |
|---|---|
| GPEL (WNA) | 1536 |
| GPEL (NRC) | 14000 |
| GPEL (ESN) | 13189 |
| DSEL (SemEval) | 3328 |
| DSEL (Blogs) | 3976 |
| DSEL (ISEAR) | 11784 |
| DSEL (Tweets) | 234923 |

- $TEC$ features extracted using baseline GPELs (WNA, NRC and ESN) (see section 3.1);
- $TEC$, $TEI$, $MEI$, $GEI$ and $GEC$ features extracted using baseline DSELs generated using PMI [22], WED [21] and sLDA [36] (see section 3.2);
- $TEC$, $TEI$, $MEI$, $GEI$ and $GEC$ features extracted using the proposed DSEL (see section 3.1). Observe that the $GEI$ and $GEC$ features are extracted using the lexicon scores for different values of threshold $\delta$. For example $GEC_{\delta 1}$ accounts only for words which have an association score with an emotion in the interval [0.25, 1]. Similarly $GEC_{\delta 2}$ and $GEC_{\delta 3}$ accounts only for words with scores in the intervals [0.5, 1] and [0.75, 1];and
- Hybrid features obtained by combining the best performing baseline features and lexicon-based features

GPELs can only be used to extract $TEC$ features, since they do not offer word-emotion quantifications needed to extract other features (refer table 1). A multi-class SVM classifier is used in all the emotion classification experiments, given its effectiveness on text classification tasks and robustness on large feature spaces. We use an optimized SVM package[7] available for all our experiments. Performance is evaluated using the standard F-score metric in all emotion classification tasks.

Word coverage of lexicons used in our comparative study for each dataset are shown in Table 4. Observe that on each data set, the coverage remains the same for all the DSELs (sLDA, PMI, WED and UMM). A higher coverage is desirable with domains like Twitter, which not only is voluminous but also contain greater numbers of unique emotive expressions compared to other domains. We expect DSELs to perform better than GPELs especially in domains (e.g. Twitter) where high lexical coverage is necessary.

### 5.3. Results and Analysis

In this section we analyse the emotion classification results obtained using baseline features, lexicon based features and a combination of them (i.e. hybrid features).

[7] https://www.csie.ntu.edu.tw/ cjlin/liblinear/

Table 5: Overall performance on different datasets with baseline features. Overall performance is measured by combining (average) the macro-averaged F-score of all the emotion classes.

| Baseline features | Overall F-Score | | | |
|---|---|---|---|---|
| | SemEval-07 | Twitter | Blogs | ISEAR |
| ngrams | 35.77 | **49.55** | **58.32** | 32.19 |
| ngrams+POS | 38.63 | 46.80 | 57.15 | 31.90 |
| ngrams+CF | 39.17 | 48.38 | 57.60 | 32.07 |
| ngrams+POS+CF | **40.99** | 47.19 | 57.03 | **32.21** |

#### 5.3.1. Performance of baseline features

Emotion classification experiments using baseline features were done incrementally by beginning with n-grams and adding one feature group (e.g. POS) at a time. Table 5 summarizes the results obtained for baseline features on the four benchmark data sets. In general, the combination of n-grams with POS features did not significantly improve emotion classification. The ineffectiveness of POS features suggests that emotions are expressed more implicitly and not just by direct words (e.g. emotional adjectives). This is similar to the findings of earlier research on emotion classification [32].

On the other hand, when n-grams are combined with contextual features performance improves over the combination of n-grams and POS features. However the combination does not consistently improve emotion classification over n-grams. This clearly suggests that the simple counts of entities such as negations, emoticons, sentiment words, punctuation etc which are found effective for sentiment classification [39] cannot be directly extended for emotion classification. Finally the combination of n-grams, POS and CF also did not consistently improve emotion classification over n-grams. These experiments clearly reflect the limitations of corpus level features identified in literature (refer section 2). In the following sections we discuss our results for lexicon based features and the hybrid features obtained by combining baseline and lexicon based features.

#### 5.3.2. Performance of lexicon based features

Emotion classification results using lexicon based features for SemEval-07, Twitter, blogs and ISEAR data sets are shown in figures 2, 3, 4 and 5 respectively. The x-axis in each of these figures indicate the different lexicon based features extracted using the knowledge of GPELs and DSELs (refer section 3.2). The y-axis indicates the overall performance (F-score) for each feature. Observe that, since GPELs are simple word-emotion lists (refer table 1), they are limited to extract only the $TEC$ feature. However in the case of DSELs performance comparison can be made across different lexicon based features extracted using the emotion quantification knowledge offered by DSELs (refer table 2).

In general features extracted from GPELs are significantly outperformed by those extracted using DSELs. The average performance improvements of all the features extracted using DSELs over those using GPELs is nearly 22%, 3% and 13% on twitter, blogs and ISEAR data sets respectively. Further the performance improvements of the proposed DSEL based features over those of the GPELs is nearly 8%, 40%, 12% and 19% on

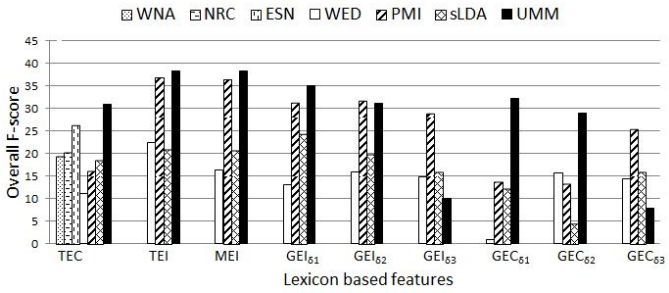


Fig. 2: Overall performance on SemEval-07 with lexicon based features. Overall performance is measured by combining (average) the macro-averaged F-score of all the emotion classes. Observe that $TEC$, $TEI$ and $MEI$ features consider all the words, whereas $GEI$ and $GEC$ features are selective. For example $GEC_{\delta1}$ accounts only for words which have an association score with an emotion in the interval [0.25, 1]. Similarly $GEC_{\delta2}$ and $GEC_{\delta3}$ accounts only for words with scores in the intervals [0.5, 1] and [0.75, 1]

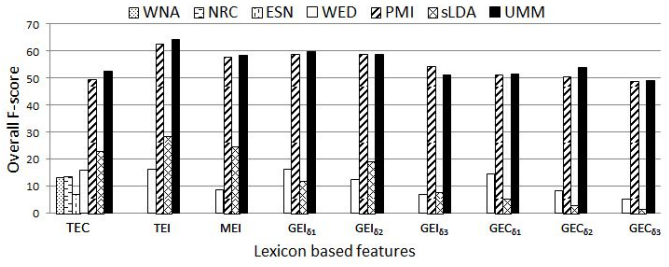


Fig. 3: Overall performance on Twitter with lexicon based features. Overall performance is measured by combining (average) the macro-averaged F-score of all the emotion classes. Observe that $TEC$, $TEI$ and $MEI$ features consider all the words, whereas $GEI$ and $GEC$ features are selective. For example $GEC_{\delta1}$ accounts only for words which have an association score with an emotion in the interval [0.25, 1]. Similarly $GEC_{\delta2}$ and $GEC_{\delta3}$ accounts only for words with scores in the intervals [0.5, 1] and [0.75, 1]

SemEval, Twitter, blogs and ISEAR data sets respectively. Essentially this confirms that GPELs are less able to capture the context in which emotions are expressed in a domain and also are less effective to model emotions in informal text streams that typically have evolving vocabularies with time.

Comparing the results in figures 2, 3, 4 and 5 suggest that $TEI$ and $MEI$ features consistently outperform $GEI$ and $GEC$ features. This is expected since the $GEI$ and $GEC$ features utilize only high intensity emotion words from a DSEL, resulting in a drop in coverage. Further a general trend of performance degradation is observed on all the data sets with $GEI$, $GEC$ features as threshold values increase from $\delta_1$ (0.25) to $\delta_2$ (0.5) to $\delta_3$ (0.75). This is expected since the proportion of high intensity emotion words, follow a decreasing series for increasing values of threshold from 0.25 to 0.75, resulting in a further drop in coverage. However it is extremely promising to note that the $GEI$ and $GEC$ features extracted from the proposed lexicon significantly outperform the $TEC$ features extracted using the GPELs. Further the proposed DSEL based features significantly outperform those extracted using WED, PMI and sLDA. In general we noticed that the generative models assumed by sLDA and WED do not effectively model the characteristics of real-world emotional data, thereby impacting the quality of the features extracted from them. Though PMI performed the best amongst the baselines, the ability of the proposed DSEL to effectively capture the associations between words and multiple emotions resulted in quality feature extraction for documents. Whilst the other DSELs also capture the word-emotion associations, the additional ability of our DSEL to discriminate between emotional and neutral words (refer table 2)improved the quality of the features extracted using its knowledge.

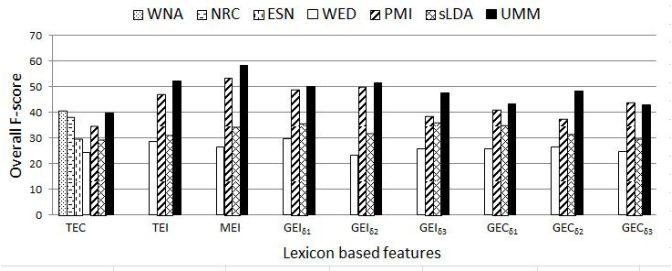


Fig. 4: Overall performance on blogs with lexicon based features. Overall performance is measured by combining (average) the macro-averaged F-score of all the emotion classes. Observe that $TEC$, $TEI$ and $MEI$ features consider all the words, whereas $GEI$ and $GEC$ features are selective. For example $GEC_{\delta1}$ accounts only for words which have an association score with an emotion in the interval [0.25, 1]. Similarly $GEC_{\delta2}$ and $GEC_{\delta3}$ accounts only for words with scores in the intervals [0.5, 1] and [0.75, 1]

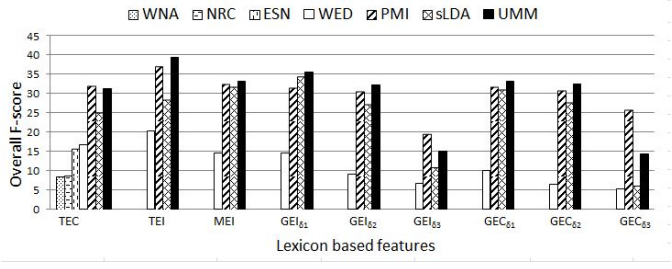


Fig. 5: Overall performance on ISEAR with lexicon based features. Overall performance is measured by combining (average) the macro-averaged F-score of all the emotion classes. Observe that $TEC$, $TEI$ and $MEI$ features consider all the words, whereas $GEI$ and $GEC$ features are selective. For example $GEC_{\delta1}$ accounts only for words which have an association score with an emotion in the interval [0.25, 1]. Similarly $GEC_{\delta2}$ and $GEC_{\delta3}$ accounts only for words with scores in the intervals [0.5, 1] and [0.75, 1]

*5.3.3. Emotion-level performance analysis*

Although the proposed DSEL in general outperformed other lexicons, we observed that the PMI lexicon is a strong competitor. Further we are also interested in comparing the performance of the lexicon based features with the baseline features discussed earlier. Accordingly we take a closer look at the baseline features[8], PMI and UMM based lexicon features by observing their performance on individual emotion classes. In particular given that not all emotions are equally complex to model, it will be useful to draw insights from those classes considered to be more challenging than others. The average F-score obtained for a class across the baseline features, lexicon based features is used as a metric to indicate its complexity. Essentially lower the F-score, the more complex (challenging) is the class prediction.

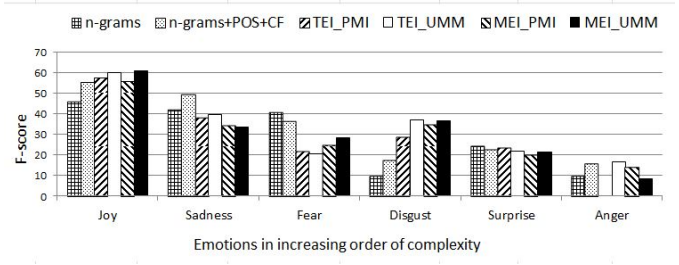


Fig. 6: Emotion-level performance of different features on SemEval-07 with respect to macro-averaged F-score. Comparative analysis is done between the best performing baseline features, best performing lexicon based features extracted using PMI and UMM.

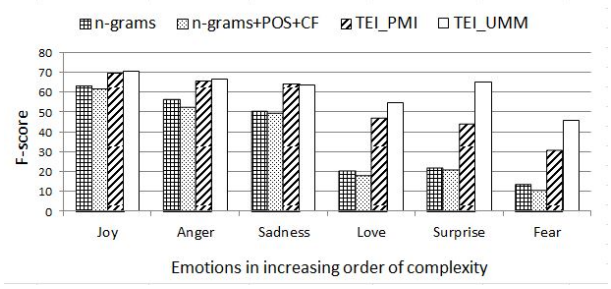


Fig. 7: Emotion-level performance of different features on Twitter with respect to macro-averaged F-score. Comparative analysis is done between the best performing baseline features, best performing lexicon based features extracted using PMI and UMM.

Figures 6, 7, 8 and 9 capture the emotion-level performance of baseline and lexicon based features. Here the x-axis plots the results in the order of increasing emotion complexity for each data set. In general the results suggest that the proposed UMM lexicon outperforms the PMI lexicon in classifying harder emotions. Similarly the proposed lexicon based features are observed to be superior to the baseline features in discriminating harder emotions on twitter and ISEAR data sets. However the performance of the proposed lexicon based features were challenged on blogs, which is explained by the skewed class distribution (see table 3) and on SemEval-07, where there is very limited data for learning lexicons (see table 3). Nevertheless the ability to have better or comparable performance to the baseline features with significantly fewer dimensions ($|E|$, where $|E|$ is the number of emotion classes in a data set) is clearly an advantage of the lexicon based feature extraction methods proposed in this paper. In the following section we discuss the results for the hybrid features obtained by combining the baseline and lexicon based features.

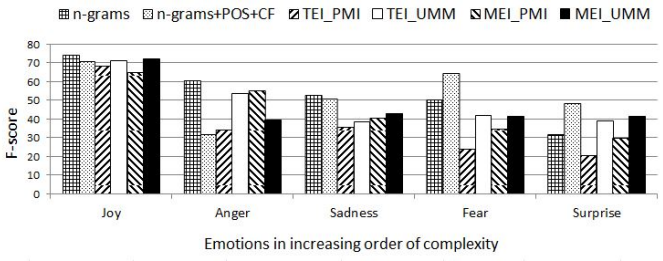


Fig. 8: Emotion-level performance of different features on Blogs with respect to macro-averaged F-score. Comparative analysis is done between the best performing baseline features, best performing lexicon based features extracted using PMI and UMM.

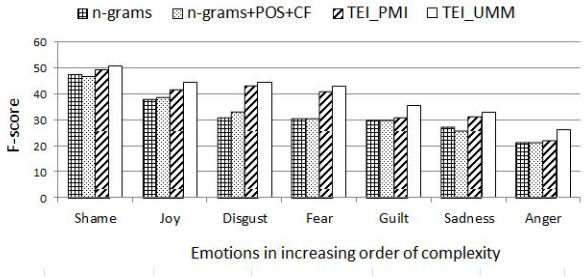


Fig. 9: Emotion-level performance of different features on ISEAR with respect to macro-averaged F-score. Comparative analysis is done between the best performing baseline features, best performing lexicon based features extracted using PMI and UMM.

*5.3.4. Performance of hybrid features*

A hybrid feature vector *hyb* is a $K + E$ dimensional feature vector obtained by combining a $K$ dimensional baseline feature vector and a $E$ dimensional lexicon based feature vector. We experimented with feature combinations of baseline[8] and lexicon based[9] features to observe for performance improvements. Emotion classification results using the hybrid features

[8] We consider the best performing baseline features for this study

[9] We consider the best performing lexicon based features derived using PMI, UMM for this study

Table 6: Emotion classification on SemEval with hybrid features. Performance is measured using macro-averaged F-score. Comparative analysis is done between systems that participated in the SemEval-07 competition, best performing baseline features, best performing lexicon based features extracted using PMI, UMM and the hybrid features.

| Features | Test set F-Score | | | | | | |
|---|---|---|---|---|---|---|---|
| | Anger | Disgust | Fear | Joy | Sadness | Surprise | Overall |
| *SemEval-07 systems* | | | | | | | |
| SWAT [45] | 7.06 | 0.0 | 18.27 | 14.91 | 17.44 | 11.78 | 11.57 |
| UA [45] | 16.03 | 0.0 | 20.06 | 4.21 | 1.76 | 15.00 | 9.51 |
| UPAR7 [45] | 3.02 | 0.0 | 4.72 | 11.87 | 17.44 | 15.00 | 8.67 |
| *Baseline features* | | | | | | | |
| (1) ngrams | 9.37 | 9.54 | **40.80** | 45.79 | 41.92 | 24.23 | 35.77 |
| (2) ngrams+POS+CF | 15.42 | 17.40 | 36.52 | 55.32 | **49.31** | 22.53 | **40.99** |
| *Lexicon based features* | | | | | | | |
| (3) $TEI_{PMI}$ | 0.00 | 28.60 | 21.53 | 57.56 | 38.34 | 24.29 | 36.78 |
| (4) $TEI_{UMM}$ | **16.78** | **36.80** | 20.63 | 59.80 | 39.69 | 21.90 | 38.16 |
| (5) $MEI_{PMI}$ | 13.86 | 34.85 | 24.67 | 56.00 | 34.32 | 20.00 | 36.54 |
| (6) $MEI_{UMM}$ | 8.30 | 36.45 | 28.13 | 61.00 | 33.63 | 21.56 | 38.23 |
| *Hybrid features* | | | | | | | |
| (1)+(3) | 8.31 | 19.00 | 28.61 | 59.64 | 37.71 | 20.00 | 37.53 |
| (1)+(4) | 5.67 | 18.82 | 33.31 | **60.00** | 31.12 | **36.40** | 38.62 |
| (1)+(5) | 7.45 | 18.21 | 28.61 | 58.71 | 38.80 | 25.70 | 38.20 |
| (1)+(6) | 15.42 | 17.41 | 36.90 | 58.61 | 40.41 | 23.21 | 39.87 |
| (2)+(3) | 5.60 | 18.21 | 23.40 | 52.90 | 30.10 | 28.60 | 33.60 |
| (2)+(4) | 8.00 | 20.00 | 32.51 | 51.83 | 29.23 | 23.00 | 33.81 |
| (2)+(5) | 12.50 | 18.80 | 27.62 | 49.72 | 35.80 | 24.00 | 34.62 |
| (2)+(6) | 12.10 | 18.20 | 32.31 | 42.30 | 35.00 | 29.30 | 33.21 |

Table 7: Emotion classification on Twitter with hybrid features. Performance is measured using macro-averaged F-score. Comparative analysis is done between best performing baseline features, best performing lexicon based features extracted using PMI, UMM and the hybrid features.

| Features | Average F-Score (10-fold cross validation) | | | | | | |
|---|---|---|---|---|---|---|---|
| | Anger | Fear | Joy | Sadness | Surprise | Love | Overall |
| *Baseline features* | | | | | | | |
| (1) ngrams | 56.68 | 13.56 | 63.34 | 50.57 | 21.65 | 20.52 | 49.55 |
| *Lexicon based features* | | | | | | | |
| (2) $TEI_{PMI}$ | 66.00 | 30.56 | 69.86 | **64.42** | 44.20 | 46.92 | 62.53 |
| (3) $TEI_{UMM}$ | **66.72** | **45.57** | **70.36** | 63.67 | **64.91** | **54.89** | **64.24** |
| *Hybrid features* | | | | | | | |
| (1)+(2) | 56.79 | 31.27 | 61.36 | 45.43 | 28.41 | 24.76 | 49.32 |
| (1)+(3) | 59.71 | 27.24 | 67.91 | 54.80 | 33.12 | 31.94 | 55.16 |

are summarized in tables 6, 7, 8 and 9. We noticed that the hybrid features involving a combination of n-grams, POS, contextual features and lexicon based features deteriorates performance. We believe this is due to the ineffective contributions of POS and contextual features as discussed earlier (refer section 5.3.1). However the hybrid features obtained by combining n-grams and lexicon based features result in performance improvements (overall F-score) over n-grams in general, except for the ISEAR data set. Further the proposed UMM lexicon derived features when combined with n-grams record significant improvements over n-grams and rest of the hybrid features. Furthermore we also noticed that the hybrid features derived using the knowledge of the proposed lexicon significantly improves performance over n-grams on complex emotions such as *surprise* on SemEval; *love, surprise* and *fear* on Twitter ; and *surprise* on blogs.

Table 8: Emotion classification on Blogs with hybrid features. Performance is measured using macro-averaged F-score. Comparative analysis is done between best performing baseline features, best performing lexicon based features extracted using PMI, UMM and the hybrid features.

| Features | Average F-Score (5-fold cross validation) | | | | | |
|---|---|---|---|---|---|---|
| | Anger | Fear | Joy | Sadness | Surprise | Overall |
| *Baseline features* | | | | | | |
| (1) ngrams | **60.30** | **50.04** | **73.92** | 52.37 | 31.32 | 58.32 |
| *Lexicon based features* | | | | | | |
| (2) $TEI_{PMI}$ | 33.94 | 23.72 | 67.92 | 35.42 | 20.14 | 47.19 |
| (3) $TEI_{UMM}$ | 53.80 | 41.70 | 71.23 | 38.50 | 38.86 | 52.18 |
| (4) $MEI_{PMI}$ | 54.90 | 34.42 | 64.50 | 40.00 | 29.32 | 53.34 |
| (5) $MEI_{UMM}$ | 39.32 | 41.63 | 72.29 | 42.68 | 41.54 | 58.16 |
| *Hybrid features* | | | | | | |
| (1)+(2) | 49.00 | 32.00 | 72.10 | 43.80 | 25.00 | 55.20 |
| (1)+(3) | 41.56 | 41.70 | 71.62 | 44.90 | 32.16 | 56.46 |
| (1)+(4) | 58.60 | 50.00 | 68.90 | 42.40 | 34.10 | 57.78 |
| (1)+(5) | 53.72 | 45.53 | 72.78 | **53.41** | **34.79** | **59.66** |

Table 9: Emotion classification on ISEAR with hybrid features. Performance is measured using macro-averaged F-score. Comparative analysis is done between best performing baseline features, best performing lexicon based features extracted using PMI, UMM and the hybrid features.

| Features | Average F-Score (5 fold cross validation) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Anger | Disgust | Fear | Guilt | Joy | Sadness | Shame | Overall |
| *Baseline features* | | | | | | | | |
| (1) ngrams | 21.11 | 31.00 | 30.62 | 29.85 | 37.86 | 27.24 | 47.56 | 32.19 |
| (2) ngrams+POS+CF | 21.12 | 33.12 | 30.40 | 29.82 | 38.71 | 25.60 | 46.74 | 32.21 |
| *Lexicon based features* | | | | | | | | |
| (3) $TEI_{PMI}$ | 21.86 | 42.92 | 40.76 | 30.93 | 41.56 | 31.30 | 49.48 | 36.96 |
| (4) $TEI_{UMM}$ | **25.96** | **44.52** | **42.88** | **35.42** | **44.45** | **32.66** | **50.54** | **39.48** |
| *Hybrid features* | | | | | | | | |
| (1)+(3) | 15.80 | 21.20 | 25.70 | 20.20 | 32.40 | 27.60 | 43.30 | 26.60 |
| (1)+(4) | 22.30 | 28.20 | 25.51 | 27.00 | 33.50 | 32.60 | 43.80 | 30.40 |
| (2)+(3) | 15.80 | 21.90 | 25.71 | 21.40 | 32.80 | 27.50 | 44.31 | 27.00 |
| (2)+(4) | 22.12 | 28.30 | 25.61 | 27.92 | 34.51 | 32.41 | 44.00 | 30.71 |

## 6. Conclusions and Future work

In this paper we extensively study the problem of emotion feature extraction using the knowledge of domain-specific lexicons (DSELs) and general purpose emotion lexicons (GPELs). We apply our unigram mixture model (UMM) based DSEL [23, 24], which is unique in its ability to quantify both emotionality and neutrality of words to extract novel features which represent documents along emotion concepts. A comparative analysis of emotion classification results on four benchmark data sets (news headlines, tweets, blogs and incident reports) suggests that the proposed features extracted using the knowledge of DSELs significantly outperform those extracted from GPELs. Further the proposed features also perform significantly better over n-gram features and their combination with features based on part-of-speech information and sentiment knowledge.

Closer examination of DSEL results show that the proposed features extracted from our UMM lexicon perform significantly better over those extracted using state-of-the-art methods such as Point-wise Mutual Information (PMI) and supervised latent dirichlet allocation (sLDA) on all the data sets. A deeper analysis of the results suggest that the proposed UMM lexicon fea-

tures are better able to classify harder emotions such as *love, fear, anger, surprise* etc. Here the use of lexicons as a means to extract new features of very low dimensions for classification purposes is shown to be a promising strategy. These findings are particularly impactful given the need for efficient and effective representations. Finally the hybrid features derived using the combination of n-grams and the proposed lexicon based features also result in consistent and significant improvements over n-gram features. This clearly confirms that a high quality lexicon which can closely capture the emotional context of a domain, when utilized effectively offers impactful knowledge for a machine learning classifier.

In future we plan to use the emotion classification system developed for analysing the emotional signatures imprinted by users in social discussion forums. We also plan to utilize the knowledge of the proposed DSEL to predict the potential evoked emotions in the readers of creative text such as fairy tales and movie plot summaries.

## References


[1] M. A. M. Shaikh, H. Prendinger, M. Ishizuka, A Linguistic Interpretation of the OCC Emotion Model for Affect Sensing from Text, Affective Information Processing, 2009, Ch. 4, pp. 45–73.

[2] B. Pang, L. Lee, Opinion mining and sentiment analysis., Foundations and Trends in Information Retrieval 2(1) (2008) 1–135.

[3] R. Calvo, S. A. D'Mello, Affect detection: An interdisciplinary review of models, methods, and their applications, IEEE Transactions on Affective Computing Volume: 1 , Issue: 1 (2010) 18– 37.

[4] H. Binali, V. Potdar, Emotion detection state of the art, in: Proc of the CUBE Int Information Technology Conference, 2012, pp. 501–507.

[5] P. Ekman, An argument for basic emotions, Cognition and Emotion 6(3) (1992) 169–200.

[6] W. G. Parrott, Emotions in social psychology, Psychology Press, Philadelphia, 2001.

[7] R. Plutchik, A general psychoevolutionary theory of emotion, In Emotion: Theory, research, and experience: Vol .1, pp 3-33.

[8] H. Gunes, M. Pantic, Automatic, dimensional and continuous emotion recognition, Int Journal of synthetic emotions, pp. 68-99.

[9] R. W. Picard, Affective Computing, 1997.

[10] F. Sebastiani, Machine learning in automated text categorization, ACM Computing Surveys 34 (2002) 1–47.

[11] H. M. Wallach, Topic modeling: beyond bag-of-words, in: Proc of the 23rd int'l conf on Machine learning, ACM, 2006, pp. 977–984.

[12] S. Scott, S. Matwin, Feature engineering for text classification, in: ICML, Vol. 99, 1999, pp. 379–388.

[13] G. V. Cormack, Feature engineering for mobile (sms) spam filtering, in: 30th ACM SIGIR Conf on Research and Development on IR, 2007.

[14] O. De Vel, A. Anderson, M. Corney, G. Mohay, Mining e-mail content for author identification forensics, ACM Sigmod 30 (4) (2001) 55–64.

[15] C. Strapparava, A. Valitutti, Wordnet-affect: an affective extension of wordnet, Tech. rep., ITC-irst, Istituto per la Ricerca Scienti?ca e Tecnologica I-38050 Povo Trento Italy (2004).

[16] S. Poria, A. Gelbukh, E. Cambria, A. Hussain, G.-B. Huang, Emosenticspace: A novel framework for affective common-sense reasoning, Knowledge-Based Systems 69, pp 108-123, 2014.

[17] S. M. Mohammad, P. Turney, Crowdsourcing a word-emotion association lexicon, Computational Intelligence, 29(3) (2013) 436–465.

[18] S. Aman, S. Szpakowicz, Identifying expressions of emotion in text, in: Proc of the 10th int conf on Text, speech and dialogue, 2007.

[19] S. M. Mohammad, Portable features for classifying emotional text, in: Conf of the NAACL: Human Language Technologies, pp 587-591,, 2012.

[20] S. Park, W. Lee, I.-C. Moon, Efficient extraction of domain specific sentiment lexicon with active learning, Elsevier Pattern Recognition Letters.

[21] Y. Rao, J. Lei, L. Wenyin, Q. Li, M. Chen, Building emotional dictionary for sentiment analysis of online news, WWW, Vol 17, pp 723-742.

[22] S. M. Mohammad, #emotional tweets, in: Proc of The First Joint Conf on Lexical and Computational Semantics, 2012.

[23] A. Bandhakavi, N. Wiratunga, P. Deepak, S. Massie, Generating a word-emotion lexicon from #emotional tweets, in: Proc of the 3rd Joint Conf on Lexical and Computational Semantics (*SEM 2014), 2014.

[24] A. Bandhakavi, N. Wiratunga, S. Massie, P. Deepak, Lexicon generation for emotion detection from text, To appear in IEEE Intelligent Systems, Jan 2017.

[25] R. Burget, J. Karasek, Z. Smekal, Recognition of emotions in czech newspaper headlines, in: Radio Engineering Vol 20 pp 39-47, 2011.

[26] G. Salton, E. A. Fox, H. Wu, Extended boolean information retrieval, ACM Communications 26 (1983) 1022–1036.

[27] B. Pang, L. Lee, S. Vaithyanathan, Thumbs up?: Sentiment classification using machine learning techniques, in: ACL-02 Conf on Empirical Methods in Natural Language Processing, Vol 10, 2002.

[28] S. Aman, S. Szpakowicz, Using roget's thesaurus for fine-grained emotion recognition, in: Int Joint Conf on Natural Language Processing, 2008.

[29] M. Purver, S. Battersby, Experimenting with distant supervision for emotion classification, in: Proc of the 13th Conf of the European Chapter of the Association for Computational Linguistics, 2012.

[30] J. D. Albornoz, L. Plaza, P. Gervas, Improving emotional intensity classification using word sense disambiguation, Special issue: Natural Language Processing and its Applications. Journal on Research in Computing Science 46 (2010) 131 – 142.

[31] D. Ghazi, D. Inkpen, S. Szpakowicz, Hierarchical approach to emotion recognition and classification in texts, in: Proc of the 23rd Canadian conference on Advances in Artificial Intelligence, 2010.

[32] W. Wang, L. Chen, K. Thirunarayan, A. P. Sheth, Harnessing twitter "big data" for automatic emotion identification, in: Proc of the ASE/IEEE Int Conf on Social Computing and Int Conf on Privacy, Security, Risk and Trust, 2012.

[33] K. Roberts, M. Roach, J. Johnson, J. Guthrie, S. M. Harabagiu, Empatweet: Annotating and detecting emotions on twitter, in: Proc. LREC, 2012, pp.3806-3813., 2012.

[34] A. Qadir, E. Riloff, Bootstrapped learning of emotion hashtags #hashtags4you, in: 4th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis (WASSA 2013), 2013.

[35] M. Yang, B. Peng, Z. Chen, a. K. C. Dingju Zhu, A topic model for building fine-grained domain-specific emotion lexicon, in: Proc of the 52nd Annual Meeting of the Assoc for Computational Linguistics,, 2014.

[36] J. D. Mcauliffe, D. M. Blei, Supervised topic models, in: Advances in Neural Information Processing Systems 20 (NIPS 2007), 2007.

[37] J. Staiano, M. Guerini, Depechemood: a lexicon for emotion analysis from crowd-annotated news, in: Proc of the 52nd Annual Meeting of the Assoc for Computational Linguistics, pp 427-433, 2014.

[38] M. Thelwall, K. Buckley, G. Paltoglou, Sentiment strength detection for the social web, Journal of the American Society for Information Science and Technology, pp 163-173, 2012.

[39] S. M. Mohammad, S. Kiritchenko, X. Zhu, Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets, in: Seventh Int Workshop on Semantic Evaluation (SemEval 2013), pp 321-327, 2013.

[40] K. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, N. A. Smith, Part of speech tagging for twitter: Annotation, features, and experiments., in: Annual Meeting Of the Assoc for Computational Linguistics, 2011.

[41] J. Suttles, N. Ide, Distant supervision for emotion classification with discrete binary values, Computational Linguistics and Intelligent Text Processing, 2013.

[42] M. Munezero, C. S. Montero, E. Sutinen, J. Pajunen, Are they different? affect, feeling, emotion,sentiment, and opinion detection in text, IEEE Transactions on Affective Computing, Vol 5 No 2, 2014.

[43] A. Esuli, S. Baccianella, F. Sebastiani, Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining, in: Proc of LREC, 2010.

[44] E. Cambria, D. Olsher, D. Rajagopal, Senticnet 3: A common and common-sense knowledge base for cognition-driven sentiment analysis, in: 28th AAAI conf on Artificial Intelligence, 2014.

[45] C. Strapparava, R. Mihalcea, Semeval-2007 task 14: Affective text, in: Proc of the 4th International Workshop on Semantic Evaluations (SemEval-2007), pp 70-74,, 2007.