Identifying Expressions of Emotion in Text

Saima Aman[1] and Stan Szpakowicz[1,2]

[1] School of Information Technology and Engineering, University of Ottawa, Ottawa, Canada
[2] Institute of Computer Science, Polish Academy of Sciences, Warszawa, Poland
{saman071, szpak}@site.uottawa.ca

Abstract. Finding emotions in text is an area of research with wide-ranging applications. We describe an emotion annotation task of identifying emotion category, emotion intensity and the words/phrases that indicate emotion in text. We introduce the annotation scheme and present results of an annotation agreement study on a corpus of blog posts. The average inter-annotator agreement on labeling a sentence as emotion or non-emotion was 0.76. The agreement on emotion categories was in the range 0.6 to 0.79; for emotion indicators, it was 0.66. Preliminary results of emotion classification experiments show the accuracy of 73.89%, significantly above the baseline.

1 Introduction

Analysis of sentiment in text can help determine the opinions and affective intent of writers, as well as their attitudes, evaluations and inclinations with respect to various topics. Previous work in sentiment analysis has been done on a variety of text genres, including product and movie reviews [9, 18], news stories, editorials and opinion articles [20], and more recently, blogs [7].

Work on sentiment analysis has typically focused on recognizing valence – positive or negative orientation. Among the less explored sentiment areas is the recognition of types of emotions and their strength or intensity. In this work, we address the task of identifying expressions of emotion in text. Emotion research has recently attracted increased attention of the NLP community – it is one of the tasks at Semeval-2007[1]; a workshop on emotional corpora was also held at LREC-2006[2].

We discuss the methodology and results of an emotion annotation task. Our goal is to investigate the expression of emotion in language through a corpus annotation study and to prepare (and place in the public domain) an annotated corpus for use in automatic emotion analysis experiments. We also explore computational techniques for emotion classification. In our experiments, we use a knowledge-based approach for automatically classifying emotional and non-emotional sentences. The results of the initial experiments show an improved performance over baseline accuracy.

The data in our experiments come from blogs. We wanted emotion-rich data, so that there would be ample examples of emotion use for analysis. Such data is

[1] http://nlp.cs.swarthmore.edu/semeval/tasks/task14/summary.shtml

[2] http://www.lrec-conf.org/lrec2006/IMG/pdf/programWSemotion-LREC2006-last1.pdf

V. Matousek and P. Mautner (Eds.): TSD 2007, LNAI 4629, pp. 196–205, 2007.

expected in personal texts, such as diaries, email, blogs and transcribed speech, and in narrative texts such as fiction. Another consideration in selecting blog text was that such text does not conform to the style of any particular genre *per se*, thus offering a variety in writing styles, choice and arrangement of words, and topics.

2 Related Work

Some researchers have studied emotion in a wider framework of *private states* [12]. Wiebe et al. [20] worked on the manual annotation of private states including emotions, opinions, and sentiment in a 10,000-sentence corpus (the MPQA corpus) of news articles. Expressions of emotions in text have also been studied within the *Appraisal Framework* [5], a functional theory of the language used for conveying attitudes, judgments and emotions [15, 19]. Neither of these frameworks deals exclusively with emotion, the focus of this paper.

In a work focused on learning specific emotions from text, Alm et al. [1] have explored automatic classification of sentences in children's fairy tales according to the basic emotions identified by Ekman [3]. The data used in their experiments was manually annotated with emotion information, and is targeted for use in a text-to-speech synthesis system for expressive rendering of stories. Read [14] has used a corpus of short stories, manually annotated with sentiment tags, in automatic emotion-based classification of sentences. These projects focus on the genre of fiction, with only sentence-level emotion annotations; they do not identify emotion indicators within a sentence, as we do in our work.

In other related work, Liu et al. [4] have utilized real-world knowledge about affect drawn from a common-sense knowledge base. They aim to understand the semantics of text to identify emotions at the sentence level. They begin with extracting from the knowledge base those sentences that contain some affective information. This information is utilized in building affective models of text, which are used to label each sentence with a six-tuple that corresponds to Ekman's six basic emotions [3]. Neviarouskaya et al. [8] have also used a rule-based method for determining Ekman's basic emotions in the sentences in blog posts.

Mihalcea and Liu [6] have focused in their work on two particular emotions – *happiness* and *sadness*. They work on blog posts which are self-annotated by the blog writers with *happy* and *sad* mood labels. Our work differs in the aim and scope from those projects: we have prepared a corpus annotated with rich emotion information that can be further used in a variety of automatic emotion analysis experiments.

3 The Emotion Annotation Task

We worked with blog posts we collected directly from the Web. First, we prepared a list of seed words for six basic emotion categories proposed by Ekman [3]. These categories represent the distinctly identifiable facial expressions of emotion – *happiness*, *sadness*, *anger*, *disgust*, *surprise* and *fear*. We took words commonly used in the context of a particular emotion. Thus, we chose “happy”, “enjoy”, “pleased” as seed words for the *happiness* category, “afraid”, “scared”, “panic” for the *fear* category, and so on. Next, using the seed words for each category, we retrieved blog posts containing one or more of those words. Table 1 gives the details of the datasets thus collected. Sample examples of annotated text appear in Table 2.

Table 1. The details of the datasets

Dataset	# posts	# sentences	Collected using seed words for
Ec-hp	34	848	*Happiness*
Ec-sd	30	884	*Sadness*
Ec-ag	26	883	*Anger*
Ec-dg	21	882	*Disgust*
Ec-sp	31	847	*Surprise*
Ec-fr	31	861	*Fear*
Total	173	5205	

Table 2. Sample examples from the annotated text

I have to look at life in her perspective, and it would break anyone’s heart. (*sadness, high*)
We stayed in a tiny mountain village called Droushia, and these people brought hospitality to incredible new heights. (*surprise, medium*)
But the rest of it came across as a really angry, drunken rant. (*anger, high*)
And I reallllly want to go to Germany – dang terrorists are making flying overseas all scary and annoying and expensive though!! (*mixed emotion, high*)
I hate it when certain people always seem to be better at me in everything they do. (*disgust, low*)
Which, to be honest, was making Brad slightly nervous. (*fear, low*)

Emotion labeling is reliable if there is more than one judgment for each label. Four judges manually annotated the corpus; each sentence was subject to two judgments. The first author of this paper produced one set of annotations, while the second set was shared by the three other judges. The annotators received no training, though they were given samples of annotated sentences to illustrate the kind of annotations required. The annotated data was prepared over a period of three months.

The annotators were required to label each sentence with the appropriate emotion category, which describes its affective content. To Ekman's six emotions [3], we added *mixed emotion* and *no emotion*, resulting in eight categories to which a sentence could be assigned. While sentiment analysis usually focuses on documents, this work’s focus is on the sentence-level analysis. The main consideration behind this decision is that there is often a dynamic progression of emotions in the narrative texts found in fiction, as well as in the conversation texts and blogs.

The initial annotation effort suggested that in many instances a sentence was found to exhibit more than one emotion – consider (1), for example, marked for both *happiness* and *surprise*. Similarly, (2) shows how more than one type of emotion can be present in a sentence that refers to the emotional states of more than one person.

(1) Everything from trying to order a baguette in the morning to asking directions or talking to cabbies, we were always pleasantly surprised at how open and welcoming they were.
(2) I felt bored and wanted to leave at intermission, but my wife was really enjoying it, so we stayed.

We also found that the emotion conveyed in some sentences could not be attributed to any basic category, for example in (3). We decided to have an additional category called *mixed emotion* to account for all such instances. All sentences that had no emotion content were to be assigned to the *no emotion* category.

(3) It's like everything everywhere is going crazy, so we don't go out any more.

In the final annotated corpus, the *no emotion* category was the most frequent. It is important to have *no emotion* sentences in the corpus, as both *positive* and *negative* examples are required to train any automatic analysis system. It should also be noted that in both sets of annotations a significant number of sentences were assigned to the *mixed emotion* category, justifying its addition in the first place.

The second kind of annotations involved assigning emotion intensity (*high, medium*, or *low*) to all emotion sentences in the corpus, irrespective the emotion category assigned to them. No intensity label was assigned to the *no emotion* sentences. A study of emotion intensity can help recognize the linguistic choices writers make to modify the strength of their expressions of emotion. The knowledge of emotion intensity can also help locate highly emotional snippets of text, which can be further analyzed to identify emotional topics. Intensity values can also help distinguish borderline cases from clear cases [20], as the latter will generally have higher intensity.

Besides labeling the emotion category and intensity, the secondary objective of the annotation task was to identify spans of text (individual words or strings of consecutive words) that convey emotional content in a sentence. We call them emotion indicators. Knowing them could help identify a broad range of affect-bearing lexical tokens and possibly, syntactic phrases. The annotators were permitted to mark in a sentence any number of emotion indicators of any length.

We considered several annotation schemes for emotion indicators. First we thought to identify only individual words for this purpose. That would simplify calculating the agreement between annotation sets. We soon realized, however, that individual words may not be sufficient. Emotion is often conveyed by longer units of text or by phrases, for example, the expressions “can't believe” and “blissfully unaware” in (4). It would also allow the study of the various linguistic features that serve to emphasize or modify emotion, as the use of word “blissfully” in (4) and “little” in (5).

(4) I can't believe this went on for so long, and we were blissfully unaware of it.
(5) The news brought them little happiness.

4 Measuring Annotation Agreement

The interpretation of sentiment information in text is highly subjective, which leads to disparity in the annotations by different judges. Difference in skills and focus of the judges, and ambiguity in the annotation guidelines and in the annotation task itself also contribute to disagreement between the judges [11]. We seek to find how much the judges agree in assigning a particular annotation by using metrics that quantify these agreements.

First we measure how much the annotators agree on classifying a sentence as an emotion sentence. Cohen's kappa [2] is popularly used to compare the extent of consensus between judges in classifying items into known mutually exclusive categories. Table 3 shows the pair-wise agreement between the annotators on emotion/non-emotion labeling of the sentences in the corpus. We report agreement values for pairs of annotators who worked on the same portion of the corpus.

Table 3. Pair-wise agreement in emotion/non-emotion labeling

	a↔b	**a↔c**	**a↔d**	**average**
Kappa	0.73	0.84	0.71	0.76

Table 4. Pair-wise agreement in emotion categories

Category	**a↔b**	**a↔c**	**a↔d**	**average**
happiness	0.76	0.84	0.71	0.77
sadness	0.68	0.79	0.56	0.68
anger	0.62	0.76	0.59	0.66
disgust	0.64	0.62	0.74	0.67
surprise	0.61	0.72	0.48	0.60
fear	0.78	0.80	0.78	0.79
mixed emotion	0.24	0.61	0.44	0.43

Within the emotion sentences, there are seven possible categories of emotion to which a sentence can be assigned. Table 4 shows the value of kappa for each of these emotion categories for each annotator pair. The agreement was found to be highest for *fear* and *happiness*. From this, we can surmise that writers express these emotions in more explicit and unambiguous terms, which makes them easy to identify. The *mixed emotion* category showed least agreement which was expected, given the fact that this category was added to account for the sentences which had more than one emotions, or which would not fit into any of the six basic emotion categories.

Agreement on emotion intensities can also be measured using kappa, as there are distinct categories – *high, medium,* and *low.* Table 5 shows the values of inter-annotator agreement in terms of kappa for each emotion intensity. The judges agreed more when the emotion intensity was high; agreement declined with decrease in the intensity of emotion. It is a major factor in disagreement that where one judge perceives a low-intensity, another judge may find no emotion.

Table 5. Pair-wise agreement in emotion intensities

Intensity	**a↔b**	**a↔c**	**a↔d**	**average**
High	0.69	0.82	0.65	0.72
Medium	0.39	0.61	0.38	0.46
Low	0.31	0.50	0.29	0.37

Emotion indicators are words or strings of words selected by annotators as marking emotion in a sentence. Since there are no predefined categories in this case, we cannot use kappa to calculate the agreement between judges. Here we need to find agreement between the sets of text spans selected by the two judges for each sentence.

Several methods of measuring agreement between sets have been proposed. For our task, we chose the measure of agreement on set-valued items (MASI), previously used for measuring agreement on co-reference annotation [10] and in the evaluation of automatic summarization [11]. MASI is a distance between sets whose value is 1 for identical sets, and 0 for disjoint sets. For sets A and B it is defined as:

MASI = J * M, where the Jaccard metric is

$$J = |A \cap B| \,/\, |A \cup B|$$

and monotonicity is

$$M = \begin{cases} 1, \mathit{if}\ A = B \\ 2/3, \mathit{if}\ A \subset B\, or\, B \subset A \\ 1/3, \mathit{if}\ A \cap B \neq \phi,\ A - B \neq \phi,\ and\ B - A \neq \phi \\ 0, \mathit{if}\ A \cap B = \phi \end{cases}$$

If one set is monotonic with respect to another, one set's elements always match those of the other set – for instance, in annotation sets {crappy} and {crappy, best} for (6). However, in non-monotonic sets, as in {crappy, relationship} and {crappy, best}, there are elements not contained in one or the other set, indicating a greater degree of disagreement. The presence of monotonicity factor in MASI therefore ensures that the latter cases are penalized more heavily than the former.

While looking for emotion indicators in a sentence, often it is likely that the judges may identify the same expression but differ in marking text span boundaries. For example in sentence (6) the emotion indicator identified by two annotators are “crappy” and “crappy relationship”, which essentially refer to the same item, but disagree on the placement of the span boundary. This leads to strings of varying lengths. To simplify the agreement measurement, we split all strings into words to ensure that members of the set are all individual words. MASI was calculated for each pair of annotations for all sentences in the corpus (see Table 6).

(6) We've both had our share of crappy relationship, and are now trying to be the best we can for each other.

We adopted yet another method of measuring agreement between emotion indicators. It is a variant of the IOB encoding [13] used in text chunking and named entity recognition tasks. We use IO encoding, in which each word in the sentence is labeled as being either In or Outside an emotion indicator text span, as shown in (7).

(7) Sorry/I for/O the/O ranting/I post/O, but/O I/O am/O just/O really/I annoyed/I.

Binary IO labeling of each word in essence reduces the task to that of word-level classification into non-emotion and emotion indicator categories. It follows that kappa can now be used for measuring agreement; pair-wise kappa values using this method are shown in Table 6. The average kappa value of 0.66 is lower than that observed at sentence level classification. This is in line with the common observation that agreement on lower levels of granularity is generally found to be lower.

Table 6. Pair-wise agreement in emotion indicators

Metric	**a↔b**	**a↔c**	**a↔d**	**average**
MASI	0.59	0.66	0.59	0.61
Kappa	0.61	0.73	0.65	0.66

5 Automatic Emotion Classification

Our long-term research goal is fine-grained automatic classification of sentences on the basis of emotion categories. The initial focus is on recognizing emotional sentences in text, regardless of their emotion category. For this experiment, we extracted all those sentences from the corpus for which there was consensus among the judges on their emotion category. This was done to form a gold standard of emotion-labeled sentences for training and evaluation of classifiers. Next, we assigned all emotion category sentences to the class “EM”, while all no emotion sentences were assigned to the class “NE”. The resulting dataset had 1466 sentences belonging to the EM class and 2800 sentences belonging to the NE class.

5.1 Feature Set

In defining the feature set for automatic classification of emotional sentences, we were looking for features which distinctly characterize emotional expressions, but are not likely to be found in the non-emotional ones. The most appropriate features that distinguish emotional and non-emotional expressions are obvious emotion words present in the sentence. To recognize such words, we used two publicly available lexical resources – the General Inquirer [16] and WordNet-Affect [17].

The General Inquirer (GI) is a useful resource for content analysis of text. It consists of words drawn from several dictionaries and grouped into various semantic categories. It lists different senses of a term and for each sense it provides several tags indicating the different semantic categories it belongs to. We were interested in the tags representing emotion-related semantic categories. The tags we found relevant are *EMOT* (emotion) – used with obvious emotion words; *Pos/Pstv* (positive) and *Neg/Ngtv* (negative) – used to indicate the valence of emotion-related words; *Intrj* (interjections); and *Pleasure* and *Pain*.

WordNet-Affect (WNA) assigns a variety of affect labels to a subset of synsets in WordNet. We utilized the publicly available lists[3] extracted from WNA, consisting of emotion-related words. There are six lists corresponding to the six basic emotion categories identified by Ekman [3].

Beyond emotion-related lexical features, we note that the emotion information in text is also expressed through the use of symbols such as emoticons and punctuation (such as "!"). We, therefore, introduced two more features to account for such symbols. All features are summarized in Table 7 (the feature vector represented counts for all features).

Table 7. Features Used in emotion classification

GI Features	WN-Affect Features	Other Features
Emotion words Positive words Negative words Interjection words Pleasure words Pain words	Happiness words Sadness words Anger words Disgust words Surprise words Fear words	Emoticons Exclamation ("!") and question ("?") marks

5.2 Experiments and Results

For our binary classification experiments, we used Naïve Bayes, and Support Vector Machines (SVM), which have been popularly used in sentiment classification tasks [6, 9]. All experiments were performed using stratified ten-fold cross validation. The naïve baseline for our experiments was 65.6%, which represents the accuracy achieved by assigning the label of the most frequent class (which in our case is NE) to all the instances in the dataset. Each sentence was represented by a 14-value vector, representing the number of occurrences of each feature type in the sentence. Table 9 shows the classification accuracy obtained with the Naïve Bayes and SVM text classifiers. The highest accuracy achieved was 73.89% using SVM, which is higher than the baseline. The improvement is statistically significant (we used the paired t-test, p=0.05).

To explore the contribution of different feature groups to the classification performance, we conducted experiments using (1) features from GI only, (2) features from WordNet-Affect only, (3) combined features from GI and WordNet-Affect, and (4) all features (including the non-lexical features). We achieved the best results when

Table 8. Emotion classification accuracy

Features	**Naïve Bayes**	**SVM**
GI	71.45%	71.33%
WN-Affect	70.16%	70.58%
GI+WN-Affect	71.7%	73.89%
ALL	**72.08%**	**73.89%**

[3] http://www.cse.unt.edu/~rada/affectivetext/data/WordNetAffectEmotionLists.tar.gz

all the features were combined. While the use of non-lexical features does not seem to affect results of SVM, it did increase the accuracy of the Naïve Bayes classifier. This suggests that a combination of features is needed to improve emotion classification results.

The results of the automatic emotion classification experiments show how external knowledge resources can be leveraged in identifying emotion-related words in text. We note, however, that lexical coverage of these resources may be limited, given the informal nature of online discourse. For instance, one of the most frequent words used for *happiness* in the corpus is the acronym "lol", which does not appear in any of these resources. In future experiments, we plan to augment the word lists obtained from GI and WordNet-Affect with such words. Furthermore, in our experiments, we have not addressed the case of typographical errors and orthographic features (for e.g. "soo sweeet") that express or emphasize emotion in text.

We also note that the use of emotion-related words is not the sole means of expressing emotion. Often a sentence, which otherwise may not have an emotional word, may become emotion-bearing depending on the context or underlying semantic meaning. Consider (8), for instance, which implicitly expresses *fear* without the use of any emotion bearing word.

(8) What if nothing goes as planned?

Therefore to be able to accurately classify emotion, we need to do contextual and semantic analysis as well.

6 Conclusion and Future Work

We address the problem of identifying expressions of emotion in text. We describe the task of annotating sentences in a blog corpus with information about emotion category and intensity, as well as emotion indicators. An annotation agreement study shows variation in agreement among judges for different emotion categories and intensity. We found the annotators to agree most in identifying instances of fear and happiness. We found that agreement on sentences with high emotion intensity surpassed that on the sentences with medium and low intensity. Finding emotion indicators in a sentence was found to be a hard task, with judges disagreeing in identifying precisely the spans of text that indicate emotion in a sentence.

We also present the results of automatic emotion classification experiments, which utilized knowledge resources in identifying emotion-bearing words in sentences. The accuracy is 73.89%, significantly higher than our baseline accuracy.

This paper described the first part of an ongoing work on the computational analysis of expressions of emotions in text. In our future work, we will use the annotated data for fine-grained classification of sentences on the basis of emotion categories and intensity. As discussed before, we plan to incorporate methods for addressing the special needs of the kind of language used in online communication. We also plan on using a corpus-driven approach in building a lexicon of emotion words. In this direction, we intend to start with the set of emotion indicators identified during the annotation process, and further extend that using similarity measures.

References

1. Alm, C.O., Roth, D., Sproat, R.: Emotions from text: machine learning for text-based emotion prediction. In: Proc. of the Joint Conf. on Human Language Technology/Empirical Methods in Natural Language Processing (HLT/EMNLP), pp. 579–586 (2005)
2. Cohen, J.: A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20, 37–46 (1960)
3. Ekman, P.: An Argument for Basic Emotions. Cognition and Emotion. 6, 169–200 (1992)
4. Liu, H., Lieberman, H., Selker, T.: A Model of Textual Affect Sensing using Real-World Knowledge. In: Proc. of the Int'l Conf. on Intelligent User Interfaces (2003)
5. Martin, J.R., White, P.R.R.: The Language of Evaluation: Appraisal in English, Palgrave, London (2005), http://grammatics.com/appraisal/
6. Mihalcea, R., Liu, H.: A corpus-based approach to finding happiness. In: The AAAI Spring Symposium on Computational Approaches to Weblogs, Stanford, CA (2006)
7. Mishne, G., Glance, N.: Predicting Movie Sales from Blogger Sentiment. In: AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs (2006)
8. Neviarouskaya, A., Prendinger, H., Ishizuka, M.: Analysis of affect expressed through the evolving language of online communication. In: Proc. of the 12th Int'l Conf. on Intelligent User Interfaces (IUI-07), Honolulu, Hawaii, pp. 278–281 (2007)
9. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment Classification using Machine Learning Techniques. In: Proc. Conf. on EMNLP (2002)
10. Passonneau, R.: Computing reliability for coreference annotation. In: Proc. International Conf. on Language Resources and Evaluation, Lisbon (2004)
11. Passonneau, R.J.: Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In: Proc. 5th Int'l Conf. on Language Resources and Evaluation (2006)
12. Quirk, R., Greenbaum, S., Leech, G., Svartvik, J.: A Comprehensive Grammar of the English Language. Longman, New York (1985)
13. Ramshaw, L.A., Marcus, M.P.: Text chunking using transformation-based learning. In: Proc. Third ACL Workshop on Very Large Corpora (1995)
14. Read, J.: Recognising affect in text using pointwise mutual information. Master's thesis, University of Sussex (2004)
15. Read, J., Hope, D., Carroll, J.: Annotating expressions of Appraisal in English. In: The Proc. of the ACL Linguistic Annotation,Workshop, Prague (2007)
16. Stone, P.J., Dunphy, D.C., Smith, M.S., Ogilvie, D.M., et al.: The General Inquirer: A Computer Approach to Content Analysis. The MIT Press, Cambridge (1966)
17. Strapparava, C., Valitutti, A.: WordNet-Affect: an affective extension of WordNet. In: Proc. 4th International Conf. on Language Resources and Evaluation, Lisbon (2004)
18. Turney, P.D.: Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In: Proc. the 40th Annual Meeting of the ACL, Philadelphia (2002)
19. Whitelaw, C., Garg, N., Argamon, S.: Using Appraisal Taxonomies for Sentiment Analysis. In: Proc. of the 2nd Midwest Comp., Linguistic Colloquium, Columbus (2005)
20. Wiebe, J., Wilson, T., Cardie, C.: Annotating expressions of opinions and emotions in language. Language Resources and Evaluation 39(2-3), 165–210 (2005)