Learning to Identify Emotions in Text

Carlo Strapparava
FBK-Irst, Italy
strappa@itc.it

Rada Mihalcea
University of North Texas
rada@cs.unt.edu

ABSTRACT

This paper describes experiments concerned with the automatic analysis of emotions in text. We describe the construction of a large data set annotated for six basic emotions: ANGER, DISGUST, FEAR, JOY, SADNESS and SURPRISE, and we propose and evaluate several knowledge-based and corpus-based methods for the automatic identification of these emotions in text.

Categories and Subject Descriptors

I.2.7 [**Natural Language Processing**]: Text analysis

General Terms

Algorithms,Experimentation

Keywords

emotion annotation, emotion analysis, sentiment analysis

1. INTRODUCTION

Emotions have been widely studied in psychology and behavior sciences, as they are an important element of human nature. They have also attracted the attention of researchers in computer science, especially in the field of human computer interaction, where studies have been carried out on facial expressions (e.g., [3]) or on the recognition of emotions through a variety of sensors (e.g., [13]).

In computational linguistics, the automatic detection of emotions in texts is becoming increasingly important from an applicative point of view. Consider for example the tasks of opinion mining and market analysis, affective computing, or natural language interfaces such as e-learning environments or educational/edutainment games.

For instance, the following represent examples of applicative scenarios in which affective analysis could make valuable and interesting contributions:

- *Sentiment Analysis.* Text categorization according to affective relevance, opinion exploration for market analysis, etc., are examples of applications of these techniques. While positive/negative valence annotation is an active area in sentiment analysis, we believe that a fine-grained emotion annotation could increase the effectiveness of these applications.

- *Computer Assisted Creativity.* The automated generation of evaluative expressions with a bias on certain polarity orientation is a key component in automatic personalized advertisement and persuasive communication.

- *Verbal Expressivity in Human Computer Interaction.* Future human-computer interaction is expected to emphasize naturalness and effectiveness, and hence the integration of models of possibly many human cognitive capabilities, including affective analysis and generation. For example, the expression of emotions by synthetic characters (e.g., embodied conversational agents) is now considered a key element for their believability. Affective words selection and understanding is crucial for realizing appropriate and expressive conversations.

This paper describes experiments concerned with the emotion analysis of news headlines. In Section 2, we describe the construction of a data set of news titles annotated for emotions, and we propose a methodology for fine-grained and coarse-grained evaluations. In Section 3, we introduce several algorithms for the automatic classification of news headlines according to a given emotion. In particular we present several algorithms, ranging from simple heuristics (e.g., directly checking specific affective lexicons) to more refined algorithms (e.g., checking similarity in a latent semantic space in which explicit representations of emotions are built, and exploiting Naïve Bayes classifiers trained on mood-labeled blogposts). Section 4 presents the evaluation of the algorithms and a comparison with the systems that participated in the SEMEVAL 2007 task on "Affective Text."

It is worth noting that the proposed methodologies are either completely unsupervised or, when supervision is used, the training data can be easily collected from online mood-annotated materials.

SAC'08 March 16-20, 2008, Fortaleza, Ceará, Brazil

2. BUILDING A DATA SET FOR EMOTION ANALYSIS

For the experiments reported in this paper we use the data set we developed for the SEMEVAL 2007 task on "Affective Text" [14].

The task was focused on the emotion classification of news headlines extracted from news web sites. Headlines typically consist of a few words and are often written by creative people with the intention to "provoke" emotions, and consequently to attract the readers' attention. These characteristics make this type of text particularly suitable for use in an automatic emotion recognition setting, as the affective/emotional features (if present) are guaranteed to appear in these short sentences.

The structure of the task was as follows:

Corpus: News titles, extracted from news web sites (such as Google news, CNN) and/or newspapers. In the case of web sites, we can easily collect a few thousand titles in a short amount of time.

Objective: Provided a predefined set of emotions (ANGER, DISGUST, FEAR, JOY, SADNESS, SURPRISE), classify the titles with the appropriate emotion label.[1]

The task was carried out in an unsupervised setting, and consequently no training was provided. The reason behind this decision is that we wanted to emphasize the study of emotion lexical semantics, and avoid biasing the participants toward simple "text categorization" approaches. Nonetheless supervised systems were not precluded from participation, and in such cases the teams were allowed to create their own supervised training sets.

Participants were free to use any resources they wanted. We provided a set of words extracted from WORDNET AFFECT [15], relevant to the six emotions of interest. However, the use of this list was entirely optional.

2.1 Data Set

The data set consisted of news headlines drawn from major newspapers such as New York Times, CNN, and BBC News, as well as from the Google News search engine. We decided to focus our attention on headlines for two main reasons. First, news have typically a high load of emotional content, as they describe major national or worldwide events, and are written in a style meant to attract the attention of the readers. Second, the structure of headlines was appropriate for our goal of conducting sentence-level annotations of emotions.

Two data sets were made available: a development data set consisting of 250 annotated headlines, and a test data set consisting of 1,000 annotated headlines.[2]

2.2 Data Annotation

To perform the annotations, we developed a Web-based annotation interface that displayed one headline at a time, together with six slide bars for emotions and one slide bar for valence. The interval for the emotion annotations was set to $[0, 100]$, where 0 means the emotion is missing from the given headline, and 100 represents maximum emotional load.

Unlike previous annotations of sentiment or subjectivity [18, 12], which typically rely on binary 0/1 annotations, we decided to use a finer-grained scale, hence allowing the annotators to select different degrees of emotional load.

The test data set was independently labeled by six annotators. The annotators were instructed to select the appropriate emotions for each headline based on the presence of words or phrases with emotional content, as well as the overall feeling invoked by the headline. Annotation examples were also provided, including examples of headlines bearing two or more emotions to illustrate the case where several emotions were jointly applicable. Finally, the annotators were encouraged to follow their "first intuition," and to use the full-range of the annotation scale bars.

2.3 Inter-Annotator Agreement

We conducted inter-tagger agreement studies for each of the six emotions. The agreement evaluations were carried out using the Pearson correlation measure, and are shown in Table 1. To measure the agreement among the six annotators, we first measured the agreement between each annotator and the average of the remaining five annotators, followed by an average over the six resulting agreement figures.

EMOTIONS	
ANGER	49.55
DISGUST	44.51
FEAR	63.81
JOY	59.91
SADNESS	68.19
SURPRISE	36.07

Table 1: Pearson correlation for inter-annotator agreement

2.4 Fine-grained and Coarse-grained Evaluations

Provided a gold-standard data set with emotion annotations, we used both fine-grained and coarse-grained evaluation metrics for the evaluation of systems for automatic emotion annotation.

Fine-grained evaluations were conducted using the Pearson measure of correlation between the system scores and the gold standard scores, averaged over all the headlines in the data set.

We also ran coarse-grained evaluations, where each emotion was mapped to a 0/1 classification (0 = [0,50), 1 = [50,100]). For the coarse-grained evaluations, we calculated precision, recall, and F-measure.

3. AUTOMATIC EMOTION ANALYSIS

3.1 Knowledge-based Emotion Annotation

We approach the task of emotion recognition by exploiting the use of words in a text, and in particular their co-occurrence with words that have explicit affective meaning. As suggested by Ortony et al. [11], we have to distinguish between words directly referring to emotional states (e.g.,

[1]The task also included a valence classification track.

[2]The data set and more information about the task can be found at the SEMEVAL 2007 web site `http://nlp.cs.swarthmore.edu/semeval`.

"fear", "cheerful") and those having only an indirect reference that depends on the context (e.g., words that indicate possible emotional causes such as "killer" or emotional responses such as "cry"). We call the former *direct affective words* and the latter *indirect affective words* [16].

As far as direct affective words are concerned, we follow the classification found in WordNet Affect.[3]. This is an extension of the WordNet database [5], including a subset of synsets suitable to represent affective concepts. In particular, one or more affective labels (*a-labels*) are assigned to a number of WordNet synsets. There are also other a-labels for those concepts representing moods, situations eliciting emotions, or emotional responses. Starting with WordNet Affect, we collected six lists of affective words by using the synsets labeled with the six emotions considered in our data set. Thus, as a baseline, we implemented a simple algorithm that checks the presence of this direct affective words in the headlines, and computes a score that reflects the frequency of the words in this affective lexicon in the text.

Sentiment analysis and the recognition of the semantic orientation of texts is an active research area in the field of natural language processing (e.g., [17, 12, 18, 9]). A crucial aspect is the availability of a mechanism for evaluating the semantic similarity among "generic" terms and affective lexical concepts. To this end we implemented a semantic similarity mechanism automatically acquired in an unsupervised way from a large corpus of texts (e.g., British National Corpus[4]). In particular we implemented a variation of Latent Semantic Analysis (LSA). LSA yields a vector space model that allows for a *homogeneous* representation (and hence comparison) of words, word sets, sentences and texts. For representing word sets and texts by means of an LSA vector, we used a variation of the *pseudo-document* methodology described in [1]. This variation takes into account also a *tf-idf* weighting schema (see [6] for more details). In practice, each document can be represented in the LSA space by summing up the normalized LSA vectors of all the terms contained in it. Thus a synset in WordNet (and even all the words labeled with a particular emotion) can be represented in the LSA space, performing the pseudo-document technique on all the words contained in the synset. In the LSA space, an emotion can be represented at least in three ways: (i) the vector of the specific word denoting the emotion (e.g. "anger), (ii) the vector representing the synset of the emotion (e.g. {`anger`, `choler`, `ire`}), and (iii) the vector of all the words in the synsets labeled with the emotion. In this paper we performed experiments with all these three representations.

Regardless of how an emotion is represented in the LSA space, we can compute a similarity measure among (generic) terms in an input text and affective categories. For example in a LSA space built form the BNC, the noun "gift" is highly related to the emotional categories JOY and SURPRISE. In summary, the vectorial representation in the LSA allows us to represent, in a *uniform* way, emotional categories, generic terms and concepts (synsets), and eventually full sentences. See [16] for more details.

[3]WordNet Affect is freely available for research purpose at `http://wndomains.itc.it` See [15] for a complete description of the resource.

[4]BNC is a very large (over 100 million words) corpus of modern English, both spoken and written (see `http://www.hcu.ox.ac.uk/bnc/`). Other more specific corpora could also be considered, to obtain a more domain oriented similarity.

3.2 Corpus-based Emotion Annotation

In addition to the experiments based on WordNet Affect, we have also conducted corpus-based experiments relying on blog entries from LiveJournal.com. We used a collection of blogposts annotated with moods that were mapped to the six emotions used in the classification. While every blog community practices a different genre of writing, LiveJournal.com blogs seem to more closely recount the goings-on of everyday life than any other blog community.

The indication of the mood is optional when posting on LiveJournal, therefore the mood-annotated posts we are using are likely to reflect the true mood of the blog authors, since they were explicitly specified without particular coercion from the interface. Our corpus consists of 8,761 blogposts, with the distribution over the six emotions shown in Table 2. This corpus is a subset of the corpus used in the experiments reported in [10].

Emotion	LiveJournal mood	Number of blogposts
ANGER	angry	951
DISGUST	disgusted	72
FEAR	scared	637
JOY	happy	4,856
SADNESS	sad	1,794
SURPRISE	surprised	451

Table 2: Blogposts and mood annotations extracted from LiveJournal

In a pre-processing step, we removed all the SGML tags and kept only the body of the blogposts, which was then passed through a tokenizer. We also kept only blogposts with a length within a range comparable to the one of the headlines, i.e. 100-400 characters. The average length of the blogposts in the final corpus is 60 words / entry. Six sample entries are shown in Table 3.

The blogposts were then used to train a Naïve Bayes classifier, where for each emotion we used the blogs associated with it as positive examples, and the blogs associated with all the other five emotions as negative examples.

4. EVALUATIONS AND RESULTS

We have implemented five different systems for emotion analysis by using the knowledge-based and corpus-based approaches described above.

1. WN-Affect presence, which is used as a baseline system, and which annotates the emotions in a text simply based on the presence of words from the WordNet Affect lexicon.

2. LSA single word, which calculates the LSA similarity between the given text and each emotion, where an emotion is represented as the vector of the specific word denoting the emotion (e.g., JOY).

3. LSA emotion synset, where in addition to the word denoting an emotion, its synonyms from the WordNet synset are also used.

ANGER
I am so angry. Nicci can't get work off for the Used's show on the 30th, and we were stuck in traffic for almost 3 hours today, preventing us from seeing them. bastards
DISGUST
It's time to snap out of this. It's time to pull things together. This is ridiculous. I'm going nowhere. I'm doing nothing.
FEAR
He might have lung cancer. It's just a rumor...but it makes sense. is very depressed and that's just the beginning of things
JOY
This week has been the best week I've had since I can't remem ber when! I have been so hyper all week, it's been awesome!!!
SADNESS
Oh and a girl from my old school got run over and died the other day which is horrible, especially as it was a very small village school so everybody knew her.
SURPRISE
Small note: French men shake your hand as they say good morning to you. This is a little shocking to us fragile Americans, who are used to waving to each other in greeting.

Table 3: Sample blogposts labeled with moods corresponding to the six emotions

4. LSA ALL EMOTION WORDS, which augments the previous set by adding the words in all the synsets labeled with a given emotion, as found in WORDNET AFFECT.

5. NB TRAINED ON BLOGS, which is a Naive Bayes classifier trained on the blog data annotated for emotions.

The five systems were evaluated on the data set of 1,000 newspaper headlines. As mentioned earlier, we conduct both fine-grained and coarse-grained evaluations. Table 4 shows the results obtained by each system for the annotation of the six emotions. The best results obtained according to each individual metric are marked in bold.

As expected, different systems have different strengths. The system based exclusively on the presence of words from the WORDNET AFFECT lexicon has the highest precision at the cost of low recall. Instead, the LSA system using all the emotion words has by far the largest recall, although the precision is significantly lower. In terms of performance for individual emotions, the system based on blogs gives the best results for JOY, which correlates with the size of the training data set (JOY had the largest number of blogposts). The blogs are also providing the best results for ANGER (which also had a relatively large number of blogposts). For all the other emotions, the best performance is obtained with the LSA models.

We also compare our results with those obtained by three systems participating in the SEMEVAL emotion annotation task: SWAT, UPAR7 and UA. Table 5 shows the results obtained by these systems on the same data set, using the same evaluation metrics. We briefly describe below each of these three systems:

UPAR7 [2] is a rule-based system using a linguistic approach. A first pass through the data "uncapitalizes" common words in the news title. The system then used the Stanford syntactic parser on the modified titles, and identifies what is being said about the main subject by exploiting the dependency graph obtained from the parser. Each word is first rated separately for each emotion and then the main subject rating is boosted. The system uses a combination of SENTIWORDNET [4] and WORDNET AFFECT [15], which

	Fine	Coarse		
	r	Prec.	Rec.	F1
ANGER				
WN-AFFECT PRESENCE	12.08	**33.33**	3.33	6.06
LSA SINGLE WORD	8.32	6.28	63.33	11.43
LSA EMOTION SYNSET	17.80	7.29	86.67	13.45
LSA ALL EMOTION WORDS	5.77	6.20	**88.33**	11.58
NB TRAINED ON BLOGS	**19.78**	13.68	21.67	**16.77**
DISGUST				
WN-AFFECT PRESENCE	-1.59	0	0	-
LSA SINGLE WORD	**13.54**	**2.41**	70.59	**4.68**
LSA EMOTION SYNSET	7.41	1.53	64.71	3.00
LSA ALL EMOTION WORDS	8.25	1.98	**94.12**	3.87
NB TRAINED ON BLOGS	4.77	0	0	-
FEAR				
WN-AFFECT PRESENCE	24.86	**100.00**	1.69	3.33
LSA SINGLE WORD	**29.56**	12.93	**96.61**	**22.80**
LSA EMOTION SYNSET	18.11	12.44	94.92	22.00
LSA ALL EMOTION WORDS	10.28	12.55	86.44	21.91
NB TRAINED ON BLOGS	7.41	16.67	3.39	5.63
JOY				
WN-AFFECT PRESENCE	10.32	**50.00**	0.56	1.10
LSA SINGLE WORD	4.92	17.81	47.22	25.88
LSA EMOTION SYNSET	6.34	19.37	72.22	30.55
LSA ALL EMOTION WORDS	7.00	18.60	**90.00**	30.83
NB TRAINED ON BLOGS	**13.81**	22.71	59.44	**32.87**
SADNESS				
WN-AFFECT PRESENCE	8.56	**33.33**	3.67	6.61
LSA SINGLE WORD	8.13	13.13	55.05	21.20
LSA EMOTION SYNSET	13.27	14.35	58.71	**23.06**
LSA ALL EMOTION WORDS	10.71	11.69	**87.16**	20.61
NB TRAINED ON BLOGS	**16.01**	20.87	22.02	21.43
SURPRISE				
WN-AFFECT PRESENCE	3.06	**13.04**	4.68	6.90
LSA SINGLE WORD	9.71	6.73	67.19	12.23
LSA EMOTION SYNSET	12.07	7.23	89.06	13.38
LSA ALL EMOTION WORDS	**12.35**	7.62	**95.31**	**14.10**
NB TRAINED ON BLOGS	3.08	8.33	1.56	2.63

Table 4: Performance of the proposed algorithms

were semi-automatically enriched on the basis of the original trial data provided during the SEMEVAL task.

UA [8] uses statistics gathered from three search engines (MyWay, AlltheWeb and Yahoo) to determine the kind and the amount of emotion in each headline. Emotion score are obtained by using Pointwise Mutual Information (PMI). First, the number of documents obtained from the three Web search engines using a query that contains all the headline words and an emotion (the words occur in an independent proximity across the Web documents) is divided by the number of documents containing only an emotion and the number of documents containing all the headline words. Second, an associative score between each content word and an emotion is estimated and used to weight the final PMI score. The final results are normalized to the 0-100 range.

SWAT [7] is a supervised system using an unigram model trained to annotate emotional content. Synonym expansion on the emotion label words is also performed, using the Roget Thesaurus. In addition to the development data provided by the task organizers, the SWAT team annotated an additional set of 1000 headlines, which was used for training.

For an overall comparison, we calculated the average over all six emotions for each system. Table 6 shows the overall results obtained by our five systems and by the three SE-

	Fine		Coarse	
	r	Prec.	Rec.	F1
ANGER				
SWAT	24.51	12.00	5.00	7.06
UA	23.20	12.74	**21.6**	16.03
UPAR7	**32.33**	**16.67**	1.66	3.02
DISGUST				
SWAT	**18.55**	0.00	0.00	-
UA	16.21	0.00	0.00	-
UPAR7	12.85	0.00	0.00	-
FEAR				
SWAT	32.52	25.00	14.40	18.27
UA	23.15	16.23	**26.27**	20.06
UPAR7	**44.92**	**33.33**	2.54	4.72
JOY				
SWAT	**26.11**	35.41	**9.44**	14.91
UA	2.35	40.00	2.22	4.21
UPAR7	22.49	**54.54**	6.66	11.87
SADNESS				
SWAT	38.98	32.50	11.92	17.44
UA	12.28	25.00	0.91	1.76
UPAR7	**40.98**	**48.97**	**22.02**	30.38
SURPRISE				
SWAT	11.82	11.86	10.93	11.78
UA	7.75	**13.70**	**16.56**	15.00
UPAR7	**16.71**	12.12	1.25	2.27

Table 5: Results of the systems participating in the the SemEval task for emotion annotations

meval systems. The best results in terms of fine-grained evaluations are obtained by the UPAR7 system, which is perhaps due to the deep syntactic analysis performed by this system. Our systems give however the best performance in terms of coarse-grained evaluations, with the WN-affect presence providing the best precision, and the LSA all emotion words leading to the highest recall and F-measure.

	Fine		Coarse	
	r	Prec.	Rec.	F1
WN-affect presence	9.54	**38.28**	1.54	4.00
LSA single word	12.36	9.88	66.72	16.37
LSA emotion synset	12.50	9.20	77.71	13.38
LSA all emotion words	9.06	9.77	**90.22**	**17.57**
NB trained on blogs	10.81	12.04	18.01	13.22
SWAT	25.41	19.46	8.61	11.57
UA	14.15	17.94	11.26	9.51
UPAR7	**28.38**	27.60	5.68	8.71

Table 6: Overall average results obtained by the five proposed systems and by the three Semeval systems

5. CONCLUSIONS

In this paper, we described experiments for the automatic annotation of emotions in text. Through comparative evaluations of several knowledge-based and corpus-based methods carried out on a large data set of 1,000 deadlines, we tried to identify the methods that work best for the annotation of emotions. In future work, we plan to explore the lexical structure of emotions, and integrate deeper semantic processing of the text into the knowledge-based and corpus-based classification methods.

Acknowledgments

Carlo Strapparava was partially supported by the HUMAINE Network of Excellence.

6. REFERENCES

[1] M. Berry. Large-scale sparse singular value computations. *International Journal of Supercomputer Applications*, 6(1):13–49, 1992.

[2] F. Chaumartin. Upar7: A knowledge-based system for headline sentiment tagging. In *Proceedings of SemEval-2007*, Prague, Czech Republic, June 2007.

[3] P. Ekman. Biological and cultural contributions to body and facial movement. In J. Blacking, editor, *Anthropology of the Body*, pages 34–84. Academic Press, London, 1977.

[4] A. Esuli and F. Sebastiani. SentiWordNet: A publicly available lexical resource for opinion mining. In *Proceedings of the 5th Conference on Language Resources and Evaluation*, Genova, IT, 2006.

[5] C. Fellbaum. *WordNet. An Electronic Lexical Database*. The MIT Press, 1998.

[6] A. Gliozzo and C. Strapparava. Domains kernels for text categorization. In *Proc. of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, Ann Arbor, June 2005.

[7] P. Katz, M. Singleton, and R. Wicentowski. Swat-mp:the semeval-2007 systems for task 5 and task 14. In *Proceedings of SemEval-2007*, Prague, Czech Republic, June 2007.

[8] Z. Kozareva, B. Navarro, S. Vazquez, and A. Montoyo. Ua-zbsa: A headline emotion classification through web information. In *Proceedings of SemEval-2007*, Prague, Czech Republic, June 2007.

[9] R. Mihalcea and H. Liu. A corpus-based approach to finding happiness. In *Proc. of Computational approaches for analysis of weblogs, AAAI Spring Symposium 2006*, Stanford, March 2006.

[10] G. Mishne. Experiments with mood classification in blog posts. In *Proceedings of the 1st Workshop on Stylistic Analysis Of Text For Information Access (Style 2005)*, Brazile, 2005.

[11] A. Ortony, G. L. Clore, and M. A. Foss. The psychological foundations of the affective lexicon. *Journal of Personality and Social Psychology*, 53:751–766, 1987.

[12] B. Pang and L. Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics*, Barcelona, Spain, July 2004.

[13] R. Picard. *Affective computing*. MIT Press, Cambridge, MA, USA, 1997.

[14] C. Strapparava and R. Mihalcea. SemEval-2007 task 14: Affective Text. In *Proceedings of SemEval-2007*, Prague, Czech Republic, June 2007.

[15] C. Strapparava and A. Valitutti. WordNet-Affect: an affective extension of WordNet. In *Proc. of 4th International Conference on Language Resources and Evaluation*, Lisbon, May 2004.

[16] C. Strapparava, A. Valitutti, and O. Stock. The affective weight of lexicon. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, Genoa, Italy, May 2006.

[17] P. Turney and M. Littman. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, 21(4):315–346, October 2003.

[18] J. Wiebe, T. Wilson, and C. Cardie. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2-3), 2005.