Portable Features for Classifying Emotional Text

Saif Mohammad
National Research Council Canada
Ottawa, Canada, K1A 0R6
saif.mohammad@nrc-cnrc.gc.ca

Abstract

Are word-level affect lexicons useful in detecting emotions at sentence level? Some prior research finds no gain over and above what is obtained with ngram features—arguably the most widely used features in text classification. Here, we experiment with two very different emotion lexicons and show that even in supervised settings, an affect lexicon can provide significant gains. We further show that while ngram features tend to be accurate, they are often unsuitable for use in new domains. On the other hand, affect lexicon features tend to generalize and produce better results than ngrams when applied to a new domain.

1 Introduction

Automatically identifying emotions expressed in text has a number of applications, including tracking customer satisfaction (Bougie et al., 2003), determining popularity of politicians and government policies (Mohammad and Yang, 2011), depression detection (Osgood and Walker, 1959; Pestian et al., 2008; Matykiewicz et al., 2009; Cherry et al., 2012), affect-based search (Mohammad, 2011), and improving human-computer interaction (Velásquez, 1997; Ravaja et al., 2006).

Supervised methods for classifying emotions expressed in a sentence tend to perform better than unsupervised ones. They use features such as unigrams and bigrams (Alm et al., 2005; Aman and Szpakowicz, 2007; Neviarouskaya et al., 2009; Chaffar and Inkpen, 2011). For example, a system can learn that the word *excruciating* tends to occur in sentences labeled with sadness, and use this word as a feature in classifying new sentences.

Approaches that do not rely on supervised training with sentence-level annotations often use affect lexicons. An affect lexicon, in its simplest form, is a list of words and associated emotions and sentiments. For example, the word *excruciating* may be associated with the emotions of sadness and fear. Note that such lexicons are at best indicators of probable emotions, and that in any given sentence, the full context may suggest that a completely different emotion is being expressed. Therefore, it is unclear how useful such word-level emotion lexicons are for detecting emotions and meanings expressed in sentences, especially since supervised systems relying on tens of thousands of unigrams and bigrams can produce results that are hard to surpass. For example, it is possible that classifiers can learn from unigram features alone that *excruciating* is associated with sadness and fear.

In this paper, we investigate whether word–emotion association lexicons can provide gains in addition to those already provided by ngram features. We conduct experiments with different affect lexicons and determine their usefulness in this extrinsic task. We also conduct experiments to determine how portable the ngram features and the emotion lexicon features are to a new domain.

2 Affect Lexicons

The WordNet Affect Lexicon (Strapparava and Valitutti, 2004) has a few thousand words annotated for associations with a number of affect categories. This includes 1536 words annotated for associations

2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 587–591,
Montréal, Canada, June 3-8, 2012.

with six emotions considered to be the most basic—joy, sadness, fear, disgust, anger, and surprise (Ekman, 1992).[1] It was created by manually identifying the emotions of a few seed words and then labeling all their WordNet synonyms with the same emotion. Affective Norms for English Words has pleasure (happy–unhappy), arousal (excited–calm), and dominance (controlled–in control) ratings for 1034 words.[2] Mohammad and Turney (2010; 2012) compiled manual annotations for eight emotions (the six of Ekman, plus trust and anticipation) as well as for positive and negative sentiment.[3] The lexicon was created by crowdsourcing to Mechanical Turk. This lexicon, referred to as the NRC word-emotion lexicon (NRC-10) version 0.91, has annotations for about 14,000 words.[4]

We evaluate the affect lexicons that have annotations for the Ekman emotions—the WordNet Affect Lexicon and the NRC-10. We also experimented with a subset of NRC-10, which we will call NRC-6, that has annotations for only the six Ekman emotions (no trust and anticipation annotations; and no positive and negative sentiment annotations).

3 Sentence Classification System

We created binary classifiers for each of the six emotions using Weka (Hall et al., 2009).[5] For example, the *Fear–NotFear* classifier determined whether a sentence expressed fear or not. We experimented with Logistic Regression (le Cessie and van Houwelingen, 1992) and Support Vector Machines (SVM). We used binary features that captured the presence or absence of unigrams and bigrams. We also used integer-valued affect features that captured the number of word tokens in a sentence associated with different affect labels in the affect lexicon being used.[6] For example, if a sentence has two joy words and one surprise word, then the joy feature has value 2, surprise has value 1, and all remaining affect labels have value 0.

[1] http://wndomains.fbk.eu/wnaffect.html

[2] http://csea.phhp.ufl.edu/media/anewmessage.html

[3] Plutchik (1985) proposed a model of 8 basic emotions.

[4] Please send an email to the author to obtain a copy of the NRC emotion lexicon. Details of the lexicon are available at: http://www.purl.org/net/saif.mohammad/research

[5] http://www.cs.waikato.ac.nz/ml/weka

[6] Normalizing by sentence length did not give better results.

emotion	# of instances	% of instances	r
anger	132	13.2	0.50
disgust	43	4.3	0.45
fear	247	24.7	0.64
joy	344	34.4	0.60
sadness	283	28.3	0.68
surprise	253	25.3	0.36
		simple average	0.54
		frequency-based average	0.43

Table 1: Inter-annotator agreement (Pearson's correlation) amongst 6 annotators on the 1000-headlines dataset.

3.1 Training and Testing within domain

As a source of labeled data for training and testing, we used the SemEval-2007 Affective Text corpus wherein newspaper headlines were labeled with the six Ekman emotions by six annotators (Strapparava and Mihalcea, 2007). For each headline–emotion pair, the annotators gave scores from 0 to 100 indicating how strongly the headline expressed the emotion. The inter-annotator agreement as determined by calculating the Pearson's product moment correlation (r) between the scores given by each annotator and the average of the other five annotators is shown in Table 1. For our experiments, we considered scores greater than 25 to indicate that the headline expresses the corresponding emotion.

The dataset was created for an unsupervised competition, and consisted of 250 sentences of trial data and 1000 sentences of test data. We will refer to them as the 250-headlines and the 1000-headlines datasets respectively. In order to use these datasets in a supervised framework, we follow Chaffar and Inkpen (2011) and report results under two settings: (1) ten-fold cross-validation on the 1000-headlines and (2) using the 1000-headlines as training data and testing on the 250-headlines dataset.

Table 2 shows results obtained by classifiers when trained on the 1000-headlines text and tested on the 250-headlines text. The rows under I give a breakdown of results obtained by the *EmotionX–NotEmotionX* classifiers when using both n-gram and NRC-10 affect features (where X is one of the six Ekman emotions). *gold* is the number of headlines expressing a particular emotion X. *right* is the number of instances that the classifier correctly

Classifier	gold	right	guess	P	R	F
I. Using affect and ngram features:						
a. NRC-10, unigrams, bigrams						
anger	66	23	55	41.8	34.8	38.0
disgust	52	8	17	47.1	15.4	23.2
fear	74	59	100	59.0	79.7	67.8
joy	77	52	102	51.0	67.5	58.1
sadness	105	71	108	65.7	67.6	66.7
surprise	43	14	67	20.9	32.6	25.4
ALL	417	227	449	50.6	54.4	**52.4**
b. NRC-6, unigrams, bigrams						
ALL	417	219	437	50.1	52.5	51.3
c. WordNet Affect, unigrams, bigrams						
ALL	417	212	490	43.3	50.8	46.7
II. Using affect features only:						
a. NRC-10						
ALL	417	282	810	34.8	67.6	46.0
b. NRC-6						
ALL	417	243	715	34.0	58.3	42.9
c. WordNet Affect						
ALL	417	409	1435	28.5	98.0	44.1
III. Using ngrams features only:						
ALL	417	210	486	43.2	50.4	46.5
IV. Random guessing:						
ALL	417	208	750	27.8	50.0	35.7

Table 2: Results on the 250-headlines dataset.

marked as expressing X. *guess* is the number of instances marked as expressing X by the classifier. Precision (P) and recall (R) are calculated as shown below:

$$P = \frac{right}{guesses} * 100 \quad (1)$$

$$R = \frac{right}{gold} * 100 \quad (2)$$

F is the balanced F-score. The ALL row shows the sums of values for all six emotions for the *gold*, *right*, and *guess* columns. The overall precision and recall are calculated by plugging these values in equations 1 and 2. Thus 52.4 is the macro-average F-score obtained by the I.a. classifiers.

I.b. and I.c. show results obtained using ngrams with NRC-6 and WordNet Affect features respectively. We do not show a breakdown of results by emotions for them and for the rows in II, III, and IV due to space constraints.

The rows in II correspond to the use of different affect features alone (no ngrams). III shows the results obtained using only ngrams, and IV shows the results obtained by a system that guesses randomly.[7]

Classifier	P	R	F
I. Using affect and ngram features:			
a. NRC-10, ngrams	44.4	61.8	**51.6**
b. NRC-6, ngrams	42.7	61.4	50.4
c. WA, ngrams	41.9	58.8	49.0
II. Using affect features only:			
a. NRC-10	24.1	95.0	38.4
b. NRC-6	24.1	95.0	38.4
c. WA	23.5	95.4	37.7
III. Using ngrams only:	42.0	59.8	49.3
IV. Random guessing:	21.7	50.0	30.3

Table 3: Cross-validation results on 1000-headlines.

Table 3 gives results obtained by cross-validation on the 1000-headlines dataset. The results in Tables 2 and 3 lead to the following observations:

- On both datasets, using the NRC-10 in addition to the ngram features gives significantly higher scores than using ngrams alone. This was not true, however, for WordNet affect.
- Using NRC-10 alone obtains almost as good scores as those obtained by the ngrams in the 250-headlines test data, even though the number of affect features (10) is much smaller than the ngram features (many thousands).
- Using annotations for all ten affect labels in NRC-10 instead of just the Ekman six gives minor improvements.
- The automatic methods perform best for classes with the high inter-annotator agreement (sadness and fear), and worst for classes with the low agreement (surprise and disgust) (Table 1).

We used the Fisher Exact Test and a confidence interval of 95% for all precision and recall significance testing reported in this paper. Experiments with support vector machines gave slightly lower F-scores than those obtained with logistic regression, but all of the above observations held true even in those experiments (we do not show those results here due to the limited space available).

[7] A system that randomly guesses whether an instance is expressing an emotion X or not will get half of the *gold* instances right. Further, it will mark half of all the instances as expressing emotion X. For ALL, $right = \frac{gold}{2}$, and $guess = \frac{instances*6}{2}$.

Emotions:	anger	3.47	joy	-0.25
	anticipn	0.08	sadness	-0.51
	disgust	0.97	surprise	-1.87
	fear	0.25	trust	0.12
Sentiment:	negative	2.38	positive	-0.31

Table 4: The coefficients of the features learned by logistic regression for the *Anger–NoAnger* classifier.

The coefficients of the features learned by the logistic regression algorithm are weights that indicate how strongly the individual features influence the decision of the classifier. The affect features of the *Anger–NoAnger* classifier learned from the 1000-sentences dataset and NRC-10 are shown in Table 4. We see that the anger feature has the highest weight and plays the biggest role in predicting whether a sentence expresses anger or not. The negative sentiment feature is also a strong indicator of anger. Similarly, the weights for other emotion classifiers were consistent with our intuition: joy had the highest weight in the *Joy–NotJoy* classifier, sadness in the *Sadness–NotSadness* classifier, and so on.

3.2 Testing on data from another domain

Hand-labeled training data is helpful for automatic classifiers, but it is usually not available for most domains. We now describe experiments to determine how well the classifiers and features cope with training on data from one source domain and testing on a new target domain. We will use the 1000-headlines dataset from the previous section as the source domain training data. As test data we will now use sentences compiled by Aman and Szpakowicz (2007) from blogs. This dataset has 4090 sentences annotated with the Ekman emotions by four annotators. The inter-annotator agreement for the different emotions ranged from 0.6 to 0.8 Cohen's kappa.

Table 5 shows the results. Observe that now the ngrams perform quite poorly; the NRC-10 affect features perform significantly better, despite each sentence being represented by only ten features. The rows in II give a breakdown of results obtained by individual *EmotionX–NotEmotionX* classifiers. Observe that the distribution of instances in this blog dataset (gold column) is different from that in the 1000-headlines (Table 1). The larger proportion of neutral instances in the blog data compared to 1000-headlines, leads to a much lower precision and F-score of the randomly-guessing classifier on the blog dataset (row IV) than on the 1000-headlines dataset.

Classifier	gold	right	guess	P	R	F
I. Using affect (NRC-10) and ngram features:						
ALL	1290	515	6717	7.7	39.9	12.9
II. Using affect (NRC-10) features only:						
anger	179	22	70	31.4	12.3	17.7
disgust	172	16	48	33.3	9.3	14.5
fear	115	32	110	29.1	27.8	28.4
joy	536	299	838	35.7	55.8	43.5
sadness	173	61	282	21.6	35.3	26.8
surprise	115	9	158	5.7	7.8	6.6
ALL	1290	439	1506	29.2	34.0	**31.4**
III. Using ngram features only:						
ALL	1290	375	7414	5.1	29.1	8.6
IV. Random guessing:						
ALL	1290	645	12270	5.3	50.0	9.6

Table 5: Results obtained on the blog dataset.

Nonetheless, the NRC-10 affect features obtain significantly higher results than the random baseline. The ngram features (row III), on the other hand, lead to scores lower than the random baseline. This suggests that they are especially domain-sensitive. Manual inspection of the regression coefficients confirms the over-fitting of ngram features. The overfitting is less for affect features, probably because of the small number of features.

4 Conclusions

Even though context plays a significant role in the meaning and emotion conveyed by a word, we showed that using word-level affect lexicons can provide significant improvements in sentence-level emotion classification—over and above those obtained by unigrams and bigrams alone. The gains provided by the lexicons may be correlated with their sizes. The NRC lexicon has fourteen times as many entries as the WordNet Affect lexicon and it gives significantly better results.

We also showed that ngram features tend to be markedly domain-specific and work well only within domains. On the other hand, affect lexicon features worked significantly better than ngram features when applied to a new domain for which there was no training data.

Acknowledgments

We thank Colin Cherry, Peter Turney, and Tara Small.

References

C. Alm, D. Roth, and R. Sproat. 2005. Emotions from text: Machine learning for text-based emotion prediction. In *Proceedings of HLT–EMNLP*, Vancouver.

S. Aman and S. Szpakowicz. 2007. Identifying expressions of emotion in text. In *Text, Speech and Dialogue*, volume 4629, pages 196–205. Springer.

J. R. G. Bougie, R. Pieters, and M. Zeelenberg. 2003. Angry customers don't come back, they get back: The experience and behavioral implications of anger and dissatisfaction in services. Open access publications from tilburg university, Tilburg University.

S. Chaffar and D. Inkpen. 2011. Using a heterogeneous dataset for emotion analysis in text. In *Canadian Conference on AI*, pages 62–67.

C. Cherry, S. M. Mohammad, and B de Bruijn. 2012. Binary classifiers and latent sequence models for emotion detection in suicide notes. *Biomedical Informatics Insights*, 5:147–154.

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

M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten. 2009. The WEKA data mining software: an update. *SIGKDD*, 11:10–18.

S. le Cessie and J. van Houwelingen. 1992. Ridge estimators in logistic regression. *Applied Statistics*, 41(1):191–201.

P. Matykiewicz, W. Duch, and J. P. Pestian. 2009. Clustering semantic spaces of suicide notes and newsgroups articles. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, BioNLP '09, pages 179–184, Stroudsburg, PA, USA. Association for Computational Linguistics.

S. M. Mohammad and P. D. Turney. 2010. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, LA, California.

S. M. Mohammad and P. D. Turney. 2012. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*.

S. M. Mohammad and T. Yang. 2011. Tracking Sentiment in Mail: How Genders Differ on Emotional Axes. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011)*, pages 70–79, Portland, Oregon. Association for Computational Linguistics.

S. M. Mohammad. 2011. From once upon a time to happily ever after: Tracking emotions in novels and fairy tales. In *Proceedings of the ACL 2011 Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, Portland, OR, USA.

A. Neviarouskaya, H. Prendinger, and M. Ishizuka. 2009. Compositionality principle in recognition of fine-grained emotions from text. In *Proceedings of ICWSM*, pages 278–281, San Jose, California.

C. E. Osgood and E. G. Walker. 1959. Motivation and language behavior: A content analysis of suicide notes. *Journal of Abnormal and Social Psychology*, 59(1):58–67.

J. P. Pestian, P. Matykiewicz, and J. Grupp-Phelan. 2008. Using natural language processing to classify suicide notes. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, BioNLP '08, pages 96–97, Stroudsburg, PA, USA. Association for Computational Linguistics.

R. Plutchik. 1985. On emotion: The chicken-and-egg problem revisited. *Motivation and Emotion*, 9(2):197–200.

N. Ravaja, T. Saari, M. Turpeinen, J. Laarni, M. Salminen, and M. Kivikangas. 2006. Spatial presence and emotions during video game playing: Does it matter with whom you play? *Presence: Teleoperators and Virtual Environments*, 15(4):381–392.

C. Strapparava and R. Mihalcea. 2007. Semeval-2007 task 14: Affective text. In *Proceedings of SemEval-2007*, pages 70–74, Prague, Czech Republic.

C. Strapparava and A. Valitutti. 2004. Wordnet-Affect: An affective extension of WordNet. In *Proceedings of LREC*, pages 1083–1086, Lisbon, Portugal.

J. D. Velásquez. 1997. Modeling emotions and other motivations in synthetic agents. In *Proceedings of the fourteenth national conference on artificial intelligence and ninth conference on Innovative applications of artificial intelligence*, AAAI'97/IAAI'97, pages 10–15. AAAI Press.