Emotion Detection in Email Customer Care

Narendra Gupta, Mazin Gilbert, and Giuseppe Di Fabbrizio
AT&T Labs - Research, Inc.
Florham Park, NJ 07932 - USA
{ngupta,mazin,pino}@research.att.com

Abstract

Prompt and knowledgeable responses to customers' emails are critical in maximizing customer satisfaction. Such emails often contain complaints about unfair treatment due to negligence, incompetence, rigid protocols, unfriendly systems, and unresponsive personnel. In this paper, we refer to these emails as *emotional emails*. They provide valuable feedback to improve contact center processes and customer care, as well as, to enhance customer retention. This paper describes a method for extracting *salient features* and identifying emotional emails in customer care. Salient features reflect customer frustration, dissatisfaction with the business, and threats to either leave, take legal action and/or report to authorities. Compared to a baseline system using word ngrams, our proposed approach with salient features resulted in a 20% absolute F-measure improvement.

1 Introduction

Emails are becoming the preferred communication channel for customer service. For customers, it is a way to avoid long hold times on call centers phone calls and to keep a record of the information exchanges with the business. For businesses, it offers an opportunity to best utilize customer service representatives by evenly distributing the work load over time, and for representatives, it allows time to research the issue and respond to the customers in a manner consistent with business policies. Businesses can further exploit the offline nature of this channel by automatically routing the emails involving critical issues to specialized representatives. Besides concerns related to products and services, businesses ensure that emails complaining about unfair treatment due to negligence, incompetence, rigid protocols and unfriendly systems, are always handled with care. Such emails, referred to as *emotional emails*, are critical to *reduce the churn* i.e., retaining customers who otherwise would have taken their business elsewhere, and, at the same time, they are a valuable source of information for improving business processes.

In recurring service oriented businesses, a large number of customer emails may contain routine complaints. While such complaints are important and are addressed by customer service representatives, our purpose here is to identify emotional emails where severity of the complaints and customer dissatisfaction are relatively high. Emotional emails may contain abusive and probably emotionally charged language, but we are mainly interested in identifying messages where, in addition to the *flames*, the customer includes a concrete description of the problem experienced with the company providing the service. In the context of customer service, customers express their concerns in many ways. Sometimes they convey a negative emotional component articulated by phrases like `disgusted` and `you suck`. In other cases, there is a minimum emotional involvement by enumerating factual sentences such as `you overcharged`, or `take my business elsewhere`. In many cases, both the emotional and factual components are actually present. In this work, we have identified eight dif-

Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, pages 10–16,
Los Angeles, California, June 2010.

ferent ways that customers use to express their emotions in emails. Throughout this paper, these ways will be referred to as *Salient Features*. We cast the identification of emotional email as a text classification problem, and show that using salient features we can significantly improve the identification accuracy. Compared to a baseline system which uses Boosting (Schapire, 1999) withnword n-grams features, our proposed system using salient features resulted in improvement in f-measure from 0.52 to 0.72.

In section 2, we provide a summary of previous work and its relationship with our contribution. In section 3, we describe our method for emotion detection and extraction of salient features. A series of experiments demonstrating improvement in classification performance is presented in section 4. We conclude the paper by highlighting the main contribution of this work in section 5.

2 Previous Work

Extensive work has been done on emotion detection. In the context of human-computer dialogs, although richer features including acoustic and intonation are available, there is a general consensus (Litman and Forbes-Riley, 2004b; Lee and Narayanan, 2005) about the use of lexical features to significantly improve the accuracy of emotion detection.

Research has also been done in predicting basic emotions (also referred to as *affects*) within text (Alm et al., 2005; Liu et al., 2003). To render speech with prosodic contour conveying the emotional content of the text, one of 6 types of human emotions (e.g., angry, disgusted, fearful, happy, sad, and surprised) are identified for each sentence in the running text. Deducing such emotions from lexical constructs is a hard problem evidenced by little agreement among humans. A $Kappa$ value of 0.24-0.51 was shown in Alm et al. (2005). Liu et al. (2003) have argued that the absence of affect laden surface features i.e., key words, from the text does not imply absence of emotions, therefore they have relied more on common-sense knowledge. Instead of deducing types emotions in each sentence, we are interested in knowing if the entire email is emotional or not. Additionally we are also interested in the intensity and the cause of those emotions.

There is also a body of work in areas of creating Semantic Orientation (SO) dictionaries (Hatzivassiloglou and McKeown, 1997; Turney and Littman, 2003; Esuli and Sebastiani, 2005) and their use in identifying emotions laden sentences and polarity (Yu and Hatzivassiloglou, 2003; Kim and Hovy, 2004; Hu and Liu, 2004) of those emotions. While such dictionaries provide a useful starting point, their use alone does not yield satisfactory results. In Wilson et al. (2005), classification of phrases containing positive, negative or neutral emotions is discussed. For this problem they show high agreement among human annotators (Kappa of 0.84). They also show that labeling phrases as positive, negative or neutral only on the basis of presence of key word from such dictionaries yields a classification accuracy of 48%. An obvious reason for this poor performance is that semantic orientations of words are context dependent.

Works reported in Wilson et al. (2005); Pang et al. (2002) and Dave et al. (2003) have attempted to mitigate this problem by using supervised methods. They report classification results using a number of different sets of features, including unigram word features. Wilson et al. (2005) reports an improvement (63% to 65.7% accuracy) in performance by using a host of features extracted from syntactic dependencies. Similarly, Gamon (2004) shows that the use of deep semantic features along with word unigrams improve performances. Pang et al. (2002) and Dave et al. (2003) on the other hand confirmed that word unigrams provide the best classification results. This is in line with our experience as well and could be due to sparseness of the data. We also used supervised methods to predict emotional emails. To train predictive models we used word ngrams (uni-, bi- and tri-grams) and a number of binary features indicating the presence of words/phrases from specific dictionaries.

Spertus (1997) discusses a system called Smoky which recognizes hostile messages and is quite similar to our work. While Smoky is interested in identifying messages that contain *flames*, our research on emotional emails looks deeper to discover the reasons for such flames. Besides word unigrams, Smoky uses rules to derive additional features for classification. These features are intended to capture different manifestations of the flames. Simi-

larly, in our work we also use rules (in our case implemented as table look-up) to derive additional features of emotional emails.

3 Emotion detection in emails

We use supervised machine learning techniques to detect emotional emails. In particular, our emotion detector is a statistical classifier model trained using hand labeled training examples. For each example, a set of salient features is extracted. The major components of our system are described below.

3.1 Classifier

For detecting emotional emails we used Boostexter as text classification. Our choice of machine learning algorithm was not strategic and we have no reason to believe that SVMs or maximum entropy–based classifiers will not perform equally well. Boostexter, which is based on the boosting family of algorithms, was first proposed by Schapire (1999). It has been applied successfully to numerous text classification applications (Gupta et al., 2005) at AT&T. Boosting builds a highly accurate classifier by combining many "weak" base classifiers, each one of which may only be moderately accurate. Boosting constructs the collection of base classifiers iteratively. On each iteration t, the boosting algorithm supplies the base learner weighted training data and the base learner generates a base classifier h_t. Set of nonnegative weights w_t encode how important it is that h_t correctly classifies each email. Generally, emails that were most often misclassified by the preceding base classifiers will be given the most weight so as to force the base learner to focus on the "hardest" examples. As described in Schapire and Singer (1999), Boostexter uses *confidence rated* base classifiers h that for every example x (in our case it is the customer emails) output a real number $h(x)$ whose sign (-1 or +1) is interpreted as a prediction(+1 indicates emotional email), and whose magnitude $|h(x)|$ is a measure of "confidence." The output of the final classifier f is $f(x) = \sum_{t=1}^{T} h_t(x)$, i.e., the sum of confidence of all classifiers h_t. The real-valued predictions of the final classifier f can be mapped onto a confidence value between 0 and 1 by a logistic function;

$$conf(x = \text{emotional email}) = \frac{1}{1 + e^{-f(x)}}.$$

The learning procedure in boosting minimizes the negative conditional log likelihood of the training data under this model, namely:

$$\sum_i \ln(1 + e^{-y_i f(x_i)}).$$

Here i iterates over all training examples and y_i is the label of ith example.

3.2 Feature extraction

Emotional emails are a reaction to perceived excessive loss of time and/or money by customers. Expressions of such reactions in emails are salient features of emotional emails. For our data we have identified the eight features listed below. While many of these features are of general nature and can be present in most customer service related emotional emails, in this paper we make no claims about their completeness.

1. Expression of negative emotions: Explicitly expressing customers affective states by phrases like `it upsets me`, `I am frustrated`;

2. Expression of negative opinions about the company: by evaluative expressions like `dishonest dealings`, `disrespectful`. These could also be insulting expressions like `stink`, `suck`, `idiots`;

3. Threats to take their business elsewhere: by expression like `business elsewhere`, `look for another provider`. These expressions are neither emotional or evaluative;

4. Threats to report to authorities: `federal agencies`, `consumer protection`. These are domain dependent names of agencies. The mere presence of such names implies customer threat;

5. Threats to take legal action: `seek retribution`, `lawsuit`. These expressions may also not be emotional or evaluative in nature;

6. Justification about why they should have been treated better. A common way to do this is

to say things like `long time customer`, `loyal customer`, etc. Semantic orientations of most phrases used to express this feature are positive;

7. Disassociate themselves from the company, by using phrases like `you people`, `your service representative`, etc. These are analogous to rule class "Noun Phrases used as Appositions" in Spertus (1997).

8. State what was done wrong to them: `grossly overcharged`, `on hold for hours`, etc. These phrases may have negative or neutral semantic orientations.

In addition to the word unigrams, salient features of emotional emails are also used for training/testing the emotional email classifier. While labeling the training data, labelers look for salient features within the email and also the severity of the loss perceived by the customer. For example, email 1 in Fig. 1 is labeled as emotional because customer perception of loss is severe to the point that the customer may cancel the service. On the other hand, email 2 is not emotional because customer perceived loss is not severe to the point of service cancellation. This customer would be satisfied in this instant if he/she receives the requested information in a timely fashion.

To extract salient features from an email, eight separate lists of phrases customers use to express each of the salient features were manually created. These lists were extracted from the training data and can be considered as basic rules that identify emotional emails. In the labeling guide for critical emails labelers were instructed to look for salient features in the email and keep a list of encountered phrases. We further enriched these lists by: a) using general knowledge of English, we added variations to existing phrases and b) searching a large body of email text (different from testing) for different phrases in which key words from known phrases participated. For example from the known phrase `lied to` we used the word `lied` and found a phrase `blatantly lied`. Using these lists we extracted eight binary salient features for each email, indicating presence/absence of phrases from the corresponding list in the email.

1. `You are making this very difficult for me. I was assured that my <SERVICE> would remain at <CURRENCY> per month. But you raised it to <CURRENCY> per month. If I had known you were going to go back on your word, I would have looked for another Internet provider. Present bill is <CURRENCY>, including <CURRENCY> for <SERVICE>.`

2. `I cannot figure out my current charges. I have called several times to straighten out a problem with my service for <PHONENO1> and <PHONENO2>. I am tired of being put on hold. I cannot get the information from the automated phone service.`

Figure 1: Email samples: 1) emotional; 2) neutral

4 Experiments and evaluation

We performed several experiments to compare the performance of our emotional email classifier with that using a ngram based text classifier. For these experiments we labeled 620 emails as training examples and 457 emails as test examples. Training examples were labeled independently by two different labelers[1] with relatively high degree of agreement among them. Kappa (Cohen, 1960) value of 0.814 was observed versus 0.5-0.7 reported for emotion labeling tasks (Alm and Sproat, 2005; Litman and Forbes-Riley, 2004a). Because of the relatively high agreement among these labelers, with different back ground, we did not feel the need to check the agreement among more than 2 labelers. Table 1 shows that emotional emails are about 12-13% of the total population.

Set	Number of examples	Critical Emails
Training	620	12%
Test	457	13%

Table 1: Distribution of emotional emails

[1]One of the labeler was one of the authors of this paper and other had linguistic back ground.

Due to the limited size of the training data we used cross validation (leave-one-out) technique on the test set to evaluate outcomes of different experiments. In this round robin approach, each example from the test set is tested using a model trained on all remaining 1076 (620 plus 456) examples. Test results on all 457 test examples are averaged.

Throughout all of our experiments, we computed the classification accuracy of detecting emotional emails using precision, recall and F-measure. Notice for our test data a classifier with majority vote has a classification accuracy of 87%, but since none of the emotional emails are identified, recall and F-measure are both zero. On the other hand, a classifier which generates many more false positives for each true positive, will have a lower classification accuracy but a higher (non-zero) F-measure than the majority vote classifier. Fig. 2 shows precision/recall curves for different experiments. The black circles represent the operating point corresponding to the best F-measure for each curve. Actual values of these points are provided in Table 2.

As a baseline experiment we used word ngram features to train a classifier model. The graph labeled as "ngram features" in Fig. 2 shows the performance of this classifier. The best F-measure in this case is only 0.52. Obviously this low performance can be attributed to the small training set and the large feature space formed by word ngrams.

	Recall	Prec.	F-Mes.
Ngram Features	0.45	0.61	0.52
Rule based: Threshholding on Salient Features counts			
≥ 4	0.41	0.93	0.57
≥ 3	0.63	0.74	0.68
≥ 2	0.81	0.53	0.63
Salient Features	0.77	0.65	0.70
ngram & Salient Features	0.65	0.81	0.72
Ngram & Random Features	0.57	0.67	0.61

Table 2: Recall and precision corresponding to best F-measure for different classifier models

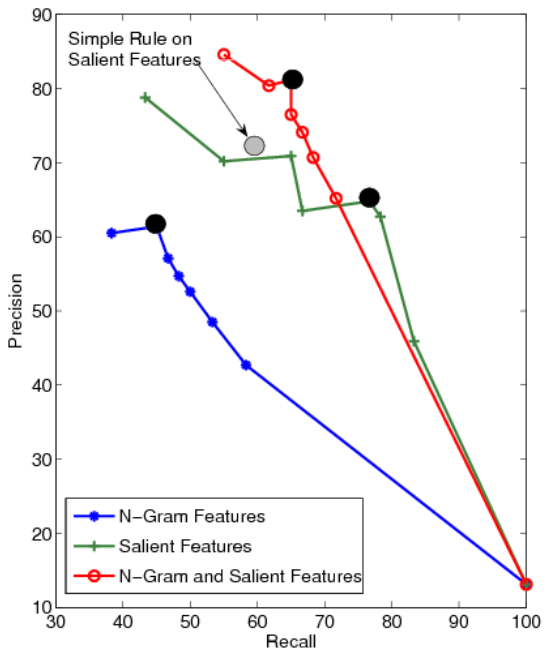

Figure 2: Precision/Recall curves for different experiments. Large black circles indicate the operating point with best F-Measure

4.1 Salient features

The baseline system was compared with a similar system using salient features. First, we used a simple classification rule that we formulated by looking at the training data. According to this rule, if an email contained three or more salient features it was classified as an emotional email. We classified the test data using this rule and obtained and an F-measure of 0.68 (see row labeled as ≥ 3 in Table 2). Since no confidence thresholding can be used with the deterministic rule, its performance is indicated by a single point marked by the gray circle in Fig. 2. This result clearly demonstrates high utility of our salient features. To verify that the salient features threshold count of 3 used in our simple classification rule is the best, we also evaluated the performance of the rule for the salient features with threshold count of 2 and 4 (row labeled as ≥ 2 and ≥ 4 in Table 2).

In our next set experiments, we trained a classifier model using salient features alone and with word ngrams. Corresponding cross validation results on the test data are annotated in Table 2 and in

Fig. 2 as "Salient Features" and "N-grams & Salient Features", respectively. Incremental improvement in best F-measure clearly shows: a) BoosTexter is able to learn better rules than the simple rule of identifying three or more salient features. b) Even though salient features provide a significant improvement in performance, there is still discriminative information in ngram features. A direct consequence of the second observation is that the detection accuracy can be further improved by extending/refining the phrase lists and/or by using more labeled data so that to exploit the discriminative information in the word ngram features.

Salient Features of emotional emails are the consequence of our knowledge of how customers react to their excessive loss. To empirically demonstrate that eight different salient features used in identification of emotional emails do provide complementary evidence, we randomly distributed the phrases in eight lists. We then used them to extract eight binary features in the same manner as before. Best F-measure for this experiment is shown in the last row of Table 2, and labeled as "N-gram & Random Features". Degradation in performance of this experiment clearly demonstrates that salient features used by us provide complimentary and not redundant information.

5 Conclusions

Customer emails complaining about unfair treatment are often emotional and are critical for businesses. They provide valuable feedback for improving business processes and coaching agents. Furthermore careful handling of such emails helps to improve customer retention. In this paper, we presented a method for emotional email identification. We introduced the notion of salient features for emotional emails, and demonstrated high agreement among two labelers in detecting emotional emails. We also demonstrated that extracting salient features from the email text and using them to train a classifier model can significantly improve identification accuracy. Compared to a baseline classifier which uses only the word ngrams features, the addition of the salient features improved the F-measure from 0.52 to 0.72. Our current research is focused on improving the salient feature extraction process. More specifically by leveraging publically available Semantic orientation dictionaries, and by enriching our dictionaries using phrases extracted from a large corpus by matching syntactic patterns of some seed phrases.

References

Alm, Cecilia and Richard Sproat. 2005. Emotional sequencing and development in fairy tales. In *Proceedings of the First International Conference on Affective Computing and Intelligent Interaction*.

Alm, Cecilia Ovesdotter, Dan Roth, and Richard Sproat. 2005. Emotions from text: machine learning for text-based emotion prediction. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Morristown, NJ, USA, pages 579–586.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20(1):37–46.

Dave, Kushal, Steve Lawrence, and David M. Pennock. 2003. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of WWW*. pages 519–528.

Esuli, A. and F. Sebastiani. 2005. Determining the semantic orientation of terms through gloss classificaion. In *Proceedings of CIKM-05, 14th ACM International Conference on Information and Knowledge Management*. Bremen, DE., pages 617–624.

Gamon, M. 2004. Sentiment classification on customer feedback data: Noisy data large feature vectors and the role of linguistic analysis. In *Proceedings of COLING 2004*. Geneva, Switzerland, pages 841–847.

Gupta, Narendra, Gokhan Tur, Dilek Hakkani-Tür, Srinivas Banglore, Giuseppe Riccardi, and Mazin Rahim. 2005. The AT&T Spoken Language Understanding System. *IEEE Transactions on Speech and Audio Processing* 14(1):213–222.

Hatzivassiloglou, Vasileios and Kathleen McKeown. 1997. Predicting the semantic orientation of ad-

jectives. In *Proceedings of the Joint ACL/EACL Conference*. pages 174–181.

Hu, Minqing and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*. pages 168–177.

Kim, Soo-Min and Eduard Hovy. 2004. Determining the sentiment of opinions. In *Proceedings of the International Conference on Computational Linguistics (COLING)*.

Lee, Chul Min and Shrikanth S. Narayanan. 2005. Toward detecting emotions in spoken dialogs. *IEEE Transactions on Speech and Audio Processing* 13(2):293–303.

Litman, D. and K. Forbes-Riley. 2004a. Annotating student emotional states in spoken tutoring dialogues. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue (SIGdial)*. Boston, MA.

Litman, D. and K. Forbes-Riley. 2004b. Predicting student emotions in computer-human tutoring dialogues. In *Proceedings of the 42nd Annual Meeting of the Association for Compuational Linguistics (ACL)*. Barcelone, Spain.

Liu, Hugo, Henry Lieberman, and Ted Selker. 2003. A model of textual affect sensing using real-world knowledge. In *IUI '03: Proceedings of the 8th international conference on Intelligent user interfaces*. ACM Press, Miami, Florida, USA, pages 125–132.

Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Philadelphia, Pennsylvania, pages 79–86.

Schapire, R.E. 1999. A brief introduction to boosting. In *Proceedings of IJCAI*.

Schapire, R.E. and Y. Singer. 1999. Improved boosting algorithms using confidence-rated predictions. *Machine Learning* 37(3):297–336.

Spertus, Ellen. 1997. Smokey: Automatic recognition of hostile messages. In *In Proc. of Innovative Applications of Artificial Intelligence*. pages 1058–1065.

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

Wilson, Theresa, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Morristown, NJ, USA, pages 347–354.

Yu, Hong and Vasileios Hatzivassiloglou. 2003. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.