A STUDY OF FINE-GRAINED SENTENCE-LEVEL EMOTION TAGGING

BY

GURMEET SINGH

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Computer Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2013

Urbana, Illinois

Adviser:

Associate Professor ChengXiang Zhai

Abstract

While there has been much work on sentiment analysis, emotion tagging has not been very well studied. Existing work has generally treated each text article as a unit for emotion tagging. In this work, we argue that it is more useful to perform emotion tagging at the sentence-level and use Conditional Random Fields (CRF) to tag sentences with five emotion tags. We propose and study multiple features, including both basic features defined on a single sentence and dependency features defined on the context of a sentence. We create two test sets, one with email messages and one with product reviews, to evaluate the proposed features. Experimental results show that in general, dependency features are beneficial, and in particular, using relative position features can significantly improve the accuracy. We also present clustering of users based on their emotional profiles as a possible application of sentence-level emotion tagging.

Acknowledgements

I would like to sincerely thank my advisor Prof. ChengXiang Zhai for his advise, encouragement and for showing me direction whenever I was lost in the research. His constant support and feedback were always there in abundance. I would also like to thank my friends Puneet Chandra and Abdel Vargas with whom I started initial work in this direction. They introduced me to the machine learning tool Weka and together we tagged and performed classification with standard classifiers.

I would also like to extend a note of thanks to my friends Manish Gupta and Hongning Wang for not only introducing me to Conditional Random Fields and the Mallet tool, but also being there to encourage me and give direction to my research. Last, but not the least, I would like to thank my family for their never ending support and motivation.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

Abbreviation	Expansion
CRF	Conditional Random Field
POS	Part of Speech
SVM	Support Vector Machine
ATF	All Transition Feature
DEF	Destination Emotions Feature
SEF	Stability for Emotions Feature
SDF	Same Destination Feature

1 Introduction

The web has seen tremendous growth of text with emotions. Users tend to express emotions in all forms ranging from emails to news article comments and product reviews. For example, the anger a customer felt about a product can be seen in this excerpt from his review *"This idea is completely wrong. Execution is not just tactics-it is a discipline and a system"*.

How to understand emotions in text is a relatively new problem that has not been well studied, yet it is an important problem as it can enable many applications. With the many opportunities available on the Internet to communicate our feelings regarding a product or an issue, companies are becoming more and more interested to know how the people are receiving the products and Government would like to know how the public is receiving or accepting their proposals, as noted in [1]. Thus, the more people talk or comment about a given topic, the more analysis would go into reasons and decision making.

On the surface, it may appear that emotion tagging is similar to sentiment tagging, which has been well studied [2], [3]. However, there are some important differences. While sentiment tagged to a text about a product gives an opinion about the user receiving the product positively or negatively [4], emotion tagging would enable to look into what the user emotionally feels about the product, its features or the company concerned. From application perspective, sentiment analysis might help a company get an insight into the popularity of a product whereas emotion tagging would help personalize the analysis by considering emotions of the user. Emotion analysis would, for instance, treat user's sad and angry emotions separately which sentiment analysis might treat as a common negative sentiment.

Emotion tagging has been studied previously in a few papers. In [5], authors tag emotions to comments on online news using meta-classification. Previous works along these lines have treated a whole article as a unit for emotion tagging. However, emotion tagging at the article level has some limitations.

The major drawback of document-based emotion tagging is that humans tend to express multiple emotions while giving their opinion. For example, when reviewing a product they are more likely to mention features that make them happy as well as those that make them sad about the product. In such a situation tagging the complete review with both Happy and Sad labels would sound ambiguous whereas tagging it as Neutral would be incorrect. More accurate and intuitive way would be to label the sentences showing happy emotions as Happy and sad ones as Sad. Such problem can be clearly seen in the following fragments of a review about a book at Amazon-

I found this book to be very well written with a rather unique approach of using... My belief is that this book will be a bit of a disappointment for those looking to find a dogmatic statement of why...

In this work, we study the problem of emotion tagging at the sentence level. That is, given an article, we attempt to tag emotion at every sentence. We consider 5 categories of emotions- *Happy, Tender, Neutral, Angry* and *Sad.* Authors of [6] and [7] have used Ekman's six basic emotions [8] for tagging- *happiness, sadness, anger, fear, surprise* and *disgust.* We believe that the set of emotions to be used depends on the application under consideration. For emotion analysis of emails, product reviews, and similar text documents the proposed set of emotions seem more appropriate as compared to *fear, surprise* and *disgust.*

Previous works on emotion tagging like [5] used techniques like meta classification of news comments using heterogeneous sources using each term as a feature in conjunction with supporting emotion tags to news articles. Some work like [6] have focused on the use of online lexical resources with naïve bayes and support vector machines [6] for emotion tagging. In this work we solve the problem of emotion tagging at sentence level with CRF. We choose to use CRF because it allows us to incorporate dependency features which, as we will show, are important to improve accuracy for sentence-level tagging.

A technical challenge in using CRF for solving our problem is the definition of features. Previous works on emotion tagging used lexical resources, unigrams, punctuations, line length, POS and emoticons as features. In our study, we will explore these basic features in the framework of CRF. In addition, we also propose new dependency features. Specifically, with CRF we have the opportunity of exploiting the context of a sentence to improve the tagging accuracy. We propose and study two kinds of such context-based features. The first is transition-based features which use the emotion label of adjacent sentences as features. Such features can be expected to capture smoothness of human transition from one emotion to another. The second feature is the relative position of the sentence in the text. This is meant to capture the part of text where the particular emotion of the user appears.

Since this work is the first to study sentence-level tagging for the given set of emotions, there doesn't exist any data set we can use for evaluation. We thus created our own data sets for emails and product reviews.

For the two datasets, CRF used with the features outperforms a naïve baseline. Evaluation results on the two data sets show several interesting findings. First, the two data sets have different behaviors. Email tagging is harder as compared to reviews. We believe this is because emails dataset has many short emails which could not exploit transition features of CRF. Also, in reviews people tend to express their opinion with more clarity. Second, the two datasets get the best prediction accuracy when classified using different set of features. We explore the reasons for that in our work. Further, it is interesting that the relative position feature is robust and can consistently improve accuracy.

The use of Conditional Random Fields for emotion tagging has been proposed in [9] for finding emotions in web blogs. In this work we make the following contributions in this direction. We introduce two new set of features to be used for emotion classification. First is that different emotion transition patterns to learn about emotional stability of users and inclination of users towards certain emotions during their communication. The second is the use of relative position of a sentence in the text which can be used to model context of sentences. Another major contribution of this work is to explore the behavior of these features for the two datasets.

Along similar lines, [6] has reported the use of CRF for word-level emotion tagging and generating sentence-level tags using weights on word-level emotion tags. This differs from our work in that we use CRF for sentence-level emotion classification instead of word-level classification to capture transitions of emotions from one sentence to another.

A closely related work is [10] where they use Conditional Random Fields to exploit the smooth transitions in sentences to do sentiment prediction. Our work differs from theirs in that we address a larger problem of emotion prediction with a set of five emotions.

MUSE [11] is an email archive browsing technique to revive memories using data mining techniques in combination with interactive visualizations. The approach they use to analyze emotions is by tracking sentimental words in the personal communication of a single user across time. To predict the emotion in an email they rely on a list of words to detect sentiments. Further, since their visualization is required to give a big picture of emotions of the user, they do not report any work on emails with mixed emotions. Previous works like SentiWordnet [12] and LIWC [13] used the similar approach. We, however, believe that while a dictionary of words will definitely improve the accuracy of any predictor, the focus for our target applications should be more on learning the pattern of emotional sentences demonstrated by users in the emails. In our work we do keep a dictionary of a limited set of words for each of the categories, but that is not the contribution of our work.

Some recent work has been reported in the literature in the direction of emotion classification. Work on Emotion Detection in Email Customer Care [14] is focused only on determining the emotional emails in the customer care scenario. They only separate emotional emails from non-emotional ones, whereas we propose to tag the sentences of a text with one of the five emotion categories.

Conditional Random Fields have been used for the problem of emotion classification before. In [15], authors use conditional random fields for sentence-level emotion classification. They use the CRFs to capture transition between different words of a sentece to enhance performance accuracy of sentence level classification. We, on the other hand, focus on capturing transition between different sentences to do the same task. This intuitively models the transition of human emotions as the person proceeds with writing a piece of text. Other works like [16] and [17] have proposed the use of CRFs for sentiment classification and opinion mining, but have not explored the features described in this work.

Another closely related work is on Emotion Tagging for Comments of Online News by Meta Classication [5] with Heterogeneous Information Sources. While this work does consider multiple emotions to be labeled in comments on news articles, they do not address the fact that a single comment can have multiple emotions. We propose tagging each sentence of the document with one of the 5 emotion labels using CRF. This is accomplished as explained in the following sections.

2 The Problem of Sentence-Level Emotion Tagging

The problem of emotion tagging differs from that of sentiment tagging in both motivation and possible applications. While sentiment tagging is more object-oriented, the problem of emotion tagging is more user or customer-oriented. The applications of the former lie in analysis of which products are more appreciated by the customers; while the latter would allow an insight into what a user feels about a product or a service and which users might need better support and direction.

Towards this end, we study the problem of emotion tagging to better analyze the emotions expressed through text. We use a set of five basic human emotions which have the potential to have applications in customer service, user categorization, corporate human resourse management and other such places. These basic emotions are Happy, Tender, Neutral, Angry and Sad. Here Neutral refers to absence of the other four emotions.

Unlike Positive/Neutral/Negative sentiments, emotions expressed in the text allow much more insight into the text by providing more features to look into. One such possibility is to exploit the natural human flow of emotions. As discussed emperically later in this work, human text shows different emotions expressed in different parts of the text. Further, there is a natural flow of emotions and there is a smooth transition of one emotion into another.

Study of text for capturing such features requires fine-course analysis of the text. To investigate the natural flow of emotions from one sentence to another, it requires the analysis to be sentence-level instead of document-level emotion tagging. Apart from providing a possibility to explore much more features, sentence level emotion tagging is better than document-level further since it allows different parts of a text document to have different emotion tags, a common occurence when a user expresses emotions through text.

For the problem of sentence-level emotion tagging, the input to the system is a text document like an email, product review, blog article or any text that is expected to have several emotions in a natural flow. The output of the system is an emotion tag assigned to each sentence in the document. We define the input to be a text t composed of n sentences as $t = \{s_1, s_2, s_3, ..., s_n\}$. The output is defined as the sequence of labels for emotions for each of the n sentences as $e = \{e_1, e_2, e_3, ..., e_n\}$, where the emotion labels are one of the five emotion categories $e_i \in \{Happy, Tender, Neutral, Angry, Sad\}$.

The motivation behind assigning a tag to each sentence rather than the complete text is that each sentence can be about different parameters of the object under consideration by the user. Hence each sentence needs to be labeled while taking into consideration the fact that human emotions do follow a natural flow from one emotion to another, as shown at the end of this section.

We use a set of five emotions for this task- Happy, Tender, Neutral, Angry and Sad. The set of emotions has been carefully selected to cover possible emotions expressed in target applications like analysis of customer emails and product reviews.

While the use of Happy, Angry and Sad seems intuitive, Tender and Neutral require justification. Tender emotion is included to cover polite sentences like greetings as well as polite disagreements. Neutral label represents lack of aforementioned emotions in the text. This category is particularly important when we consider transition in emotions from one sentence to another. For instance, an Angry sentence is not expected to transition to a Happy one directly, but follows a smooth transition through Neutral sentences in between.

The following email with 5 sentences clearly shows a smooth transition from Sad to Happy through Neutral-

Again, please accept my apology. I just need to move forward and stop reflecting on the past. It is so difficult when you have given your entire career for an organization that still doesn't see value in you. But, that is my problem not yours. In regards to your party, your right, I laughed at the invitation.

3 Conditional Random Fields for Emotion Tagging

Tagging of emotions at sentence level is a difficult task because it needs automated understanding of human sentences which are often ambiguous and depend on the context in which they are said. Also, sentences can have very few words which makes their classification even more difficult. Hence, it becomes necessary to take into account the context in which a sentence was said.

Most classifiers used in the existing work [18] have the limitation that they do not consider the transition between labels generated. This use of transitions between labels as a feature is of particular importance in the situations when either the label of one sentence depends on another or when we can leverage on the use of labels of neighboring sentences to have a better estimate for label of the sentence in hand.

In order to model the emotion of a sentence, it is necessary to consider the context. We capture the context of a sentence in two ways. First we consider the smooth flow of emotions from one sentence to another. This is necessary because human emotions tend to follow smooth transitions from one sentence to another. It is uncommon to observe a dramatic changes from emotion to another as discussed in Section 3.2.2. Second is the use of position of the sentence in the text as a feature.

To capture the transitions between emotion labels, we propose to use CRF which provides a principled way to capture such dependencies. Below we first give a brief introduction to CRF followed by details of how emotion transitions and position of sentences can be incorporated as features for emotion prediction.

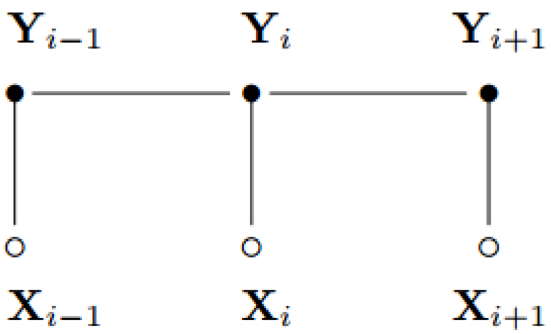


Figure 3.1: Chained-structured case of conditional random fields.

3.1 Conditional Random Fields

Conditional Random Fields [19] take an input X as random variable overa sequence to be labeled and generate an output random variable Y over corresponding label sequences. The random variables (X, Y) create a conditional random field if Y_i conditioned on X follow the Markov property with respect to the graph made by labels. That is, $p(Y_v|X, Y_w, w \neq v) = p(Y_v|X, Y_w, w \sim v)$, where $w \sim v$ means w and v are neighbors in the graph G [19].

In our scenario, X_is are the sentences of a single text. The corresponding output labels $Y_i \in \{Happy, Tender, Neutral, Angry, Sad\}$. The graph G is a linear chain where each sentence corresponds to a node in graph. X and Y are indexed by vertices in the graph G. Following the Markov property, the label of one sentence Y_v depends only on the input text X and output labels of sentences neighboring to Y_v.

For labeling the text with emotion tags, we are interested in conditional probability of the label Y_i given the sentence being labeled X_i. Using a probabilistic model, we plan to capture the emotion transitions between different sentences of the text. Conditional random fields allow us to directly capture dependency features of the observed sequence.

Consider the chain structure shown in the Figure 3.1 that shows a typical chained-structured case of CRFs for sequences. In our study, X_is stand for the sentences and Y_is stand for the label sequence. The figure illustrates that the dependencies of Y conditioned on X form a chain.

3.2 Feature Set

We now discuss how to define the features to be used in our sentence-level emotion tagging system.

3.2.1 Basic Features

Our starting point is the basic features one can define on a sentence. In the previous works, standard features like unigrams, line length, punctuations, capitalizations and removal of stop words have been used. We thus also include these as our basic features. Some work has also been focused on using emotion lexicon to improve accuracy [18]. We believe that including an extended lexicon of emotional words would definitely improve the accuracy of the classifier; however that is not the focus of our work. Hence we use a limited word dictionary to augment the features.

3.2.2 Transitions Between Emotions

We consider the flow of emotions from one sentence to another as a feature for CRF. This is necessary because human emotions tend to follow smooth transitions from one sentence to another. This is captured in CRF using transitions between adjacent emotion labels.

Let us define E as the set of five emotions. Let e_i and e_{i+1} be two adjacent emotion tags for the sentences s_i and s_{i+1} of a text t. In this work, we call e_i as the source emotion and e_{i+1} as the destination emotion. Let T denote the transition from source emotion e_i to the destination emotion e_{i+1} as $T(e_i, e_{i+1})$. Each emotion transition feature can be defined as a function of source and destination emotions $f(e_i, e_{i+1})$. We consider these transitions in the following ways to capture different characteristics of human emotional behavior.

- **Type 1: No Transitions** When CRF is used in this configuration, we do not use the emotion tags of adjacent sentences of a current sentence as a feature for emotion classification of the current sentence. This is used as a baseline to evaluate the effect of transition features described ahead.

- **Type 2: All Transitions Feature** This feature set considers every transition of emotion as a feature. The feature function $ATF(e_i, e_{i+1})$ will have value 1 when the transition is from sentence with label e_i to sentence with label e_{i+1}. The 5x5 emotion transitions result in 25 features.

- **Type 3: Destination Emotions Feature** This feature set considers the transitions of emotions which have emotion e_i as the destination emotion. The feature function $DEF(e_i) = 1$ only for the set of emotion transitions that have e_i as the destination emotion, that is for the transition $T(e_j, e_i) \in \{e_j \rightarrow e_i \forall e_j \in E\}$. Such features represent the likelihood of the user to switch from any emotion to a particular destination emotion. They capture the inclination of a user for a particular emotion.

- **Type 4: Stability for Emotions Feature** This feature set captures the stability a user has for a particular emotion. The feature function $SEF(e_i) = 1$ iff the transition of emotion is $T(e_i, e_i) = \{(e_i \rightarrow e_i)\}$. These features represent how likely is the user to stay in a particular emotion and how likely he might be to change it.

- **Type 5: Same Destination Feature** This feature considers if the user has the destination emotion same as the source emotion for the transition. The

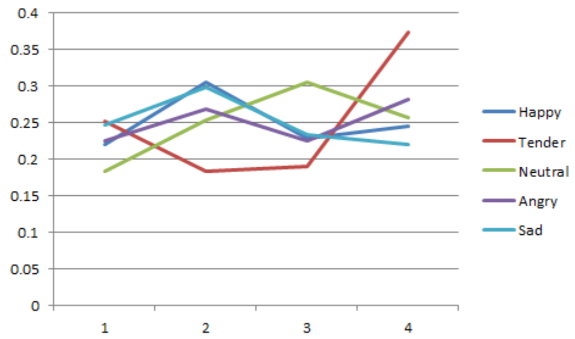


Figure 3.2: Distribution of emotions in the four quarters of the emails.

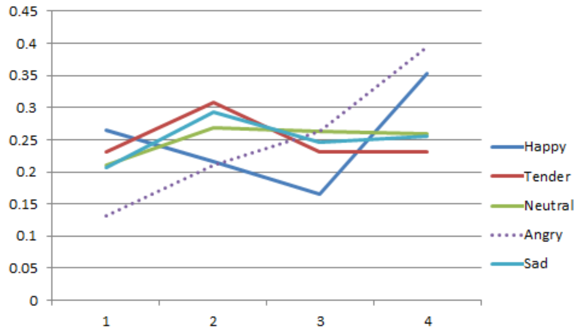


Figure 3.3: Distribution of emotions in the four quarters of the reviews.

feature function $SDF(e_i, e_{i+1}) = 1$ iff the emotion transition is $T(e_i, e_{i+1}) \in \{e_i \rightarrow e_{i+1}, e_{i+1} = e_i \forall e_i \in E\}$. Since these features capture whether or not the user would change the present emotion, they represent the overall stability of the user for all emotions.

3.2.3 Relative Position of Sentence

Another relevant feature evaluated is the relative position of a sentence in the text. In general, it has been observed that users tend to express different emotions in different parts of text. We found empirically that people end the email on a more polite note as shown in Figure 3.2 and express their sadness towards the end of product review as shown in Figure 3.3.

For instance, the following email shows how a user tends to start with *happy* emotion and goes on to communicate *neutral* sentences towards the end in accordance with Figure 3.2.

Great! Looks like the 22nd and 23rd will work turns out I'll be in Chicago around those days. What is the name of the B&B? I'll call for reservations. who do you want to come along? It will be great to see you guys again! I have a trek July 10-16 with 28 riders. So that week is definitely out?

The same user in a different email expresses the *sad* emotion only after a set of preceding *neutral* sentences.

I presume we're talking about a weekend. How many will be coming and for how many days?? I have bought 9 new horses this past month. want to try some of them out? Russia is now a gelding but I bought two Arabian stallions. We'll see which (if any) of them remain so. We no longer routinely provide lunches on all day rides, I'm sorry about the bad notice.

To capture this observation for emotion prediction, we consider the relative position of a sentence in the text as another feature for the classifier. For every review and email, we divide each text document into k equal sized partitions. For every sentence in the text we add a new feature for it that is the partition number the sentence belongs to within the text. Absence of this feature corresponds to the situation with $k = 1$ and dividing document to four quarters means $k = 4$.

4 Evaluation Results

Our evaluation procedure studies which features are the most useful for sentence-level emotion tagging using CRF. We also study if the two datasets behave similarly for the problem of sentence-level emotion classification. The implementation of CRF features is done on top of the tool Mallet. Mallet is an open source tool for machine learning applications to the text [20].

4.1 Data Sets

No previous work addressed our problem exactly, thus no existing data set can be used. We created new data sets in order to incorporate the labels for the five emotions- *Happy, Tender, Neutral, Angry, Sad.*

We use two public datasets- one for emails and other for product reviews. For emails we used the Enron dataset [21] and for reviews we used the Amazon reviews dataset [22]. From each of them we randomly select a subset of users and randomly select their emails or reviews to be labeled, to avoid introduction of any bias.

The data is manually labeled by three human annotators. The set of documents to be labeled was randomly divided into three subsets and each annotator was given two of these subsets. For further experiments, we only considered the documents for which the labels of the two annotators matched. This led to discarding a total of 42 annotated emails covering 187 sentences and 14 reviews covering 183 sentences from the labeled datasets.

The email dataset used further has 1197 labeled sentences covering 334 emails.

Label	Emails Dataset	Reviews Dataset
Happy	17.42	22.32
Tender	23.28	8.39
Neutral	37.78	56.00
Angry	7.12	3.50
Sad	14.40	9.78

Table 4.1: Percentage of sentences with each category of labels for the email and review datasets

The reviews dataset has 1054 labeled sentences covering 82 reviews. The email dataset has on an average 3.57 sentences with standard deviation of 4.44 and the reviews have an average length of 13.2 sentences with 8.69 as the standard deviation.The distribution of labels in the two datasets is as shown in Table 4.1.

4.2 Baseline

In this work, we focus on studying the effect of proposed features for emotion prediction using CRF. For this, the baseline for our experiments is the accuracy achieved by assigning the label of the most frequent class to all the instances in the dataset. For both the datasets the most frequent class is *Neutral*. Assigning this label to all instances of email dataset gives an accuracy of 37.77%. Similarly, for reviews dataset the baseline accuracy is 56%.

Further, we evaluate the effect of these features on standard classifiers for the problem of sentence-level emotion tagging. Several researchers [9], have used well-known classifiers like Naïve Bayes and Support Vector Machines with aforementioned basic features. We also perform the experiments using Decision Tree, Ada Boost and Random Forests to evaluate their accuracy in tagging emotions when used with the proposed features for the two data sets.

4.3 Performance Measures

We compare the performance of classifier with the proposed features using prediction accuracy. In addition to that, we also report per-category precision, recall and F1 values for a complete insight into performance of the features.

4.4 Experiment Design

The labeled dataset is divided into 10 partitions and 10-fold cross validation is performed using 9 partitions as training dataset and 1 partition as the test dataset. The input to the classifier is the set of documents with labeled sentences for training and the test data sets. Classifier learns feature weights using the training set and report the accuracy of prediction for the test set to give an overall score of prediction accuracy for the classifier.

There are two primary questions to be answered. First is whether the use of emotion transitions and relative positions of sentence as features improves the performance of the classifier. Second is to study how the two data sets differ with regard to emotion classification.

We perform a series of tests to find out performance of proposed features with CRF. We also report the findings of tests performed to find out the best possible set of emotion-transition features when training Conditional Random Fields for classification. Followed by this, we study the effect of the use of relative position of sentences in the text as a feature for emotion classification. We report the results in the following subsection.

4.5 Results

The experiments to study the emotion classification were performed in two stages. In the first part of the experiments we study the effect of the new features introduced for Conditional Random Fields. Here we compare the prediction accuracy with a naïve baseline of assigning the most popular emotion tag to all sentences. From these experiments we find out the set of features which work the best for both the datasets. We use these features and compare the accuracy of CRF with the standard classifiers- Naïve Bayes, Decision Tree, Ada Boost, Random Forests and Support Vector Machines.

4.5.1 Effect of New Features

We conducted experiments to analyze the effect emotion transitions and relative position of sentence independently, and collectively. First we test the accuracy of CRF using the different kinds of transition features for both the datasets and compare that to the baseline classifiers. For this part of the experiment, we do not use relative position of the sentences as a feature. Next we introduce relative position features by dividing the text documents into k partitions. Values of k tested for the two datasets are $k = 2, 3, 4$ and 5. The value of $k = 1$ corresponds to the case when relative position features have not been used and entire text document is a single collection.

Emotion Transition Features

First part of our experiments was focused on finding the best possible emotion transition features for Conditional Random Fields for the two datasets. For this we tried five types of features as listed in Section 3.2.2. The results are reported in Table 4.2 for emails in first column and for reviews in second column. Note that

Features	Emails Accuracy	Reviews Accuracy
CRF Type 1	38.23	55.07
CRF Type 2	46.37	54.14
CRF Type 3	47.42	53.65
CRF Type 4	33.16	54.64
CRF Type 5	38.51	50.24
Naïve Baseline	37.77	56

Table 4.2: Percentage accuracy of prediction for different types of emotion-transition features used for CRF

Type 1 features for CRF means transition features were not considered for that system. It is evident that Type 2 and Type 3 features are the most accurate in predicting emotions across the two datasets.

As seen from the table, introduction of Type 2 and Type 3 features give accuracy significantly higher than that without transition features (Type 1). With these two, CRF is able to outperform the baseline 37.77%. On the other hand, for reviews the prediction accuracy with only transition features does not outperform baseline 56%. Introduction of relative position labels as features improves the performance for the two datasets.

Relative Postion of Sentences

To study the effect of relative position of a sentence in the text as a feature for emotion prediction, we performed experiments to compare the effect of partitioning each text documents into k equal parts, with $k = 2, 3, 4, 5$ and observing what value of k gives highest accuracy for the two datasets.

Table 4.3 and Table 4.4 show a comparison of classifier accuracy for the different values of k. It is observed that $k = 5$ works the best for email dataset and $k = 4$ for reviews.

Two major observations can be made from these results. First is that even with-

Classifier	k=1	k=2	k=3	k=4	k=5
CRF Type 1	38.23	43.76	42.68	42.86	42.48
CRF Type 2	46.37	42.37	43.22	43.69	45.1
CRF Type 3	47.42	47.99	47.61	48.07	48.38
CRF Type 4	33.16	43.54	42.32	41.52	41.39
CRF Type 5	38.51	38.19	36.97	38.97	37.29
Naïve Baseline	37.77	37.77	37.77	37.77	37.77

Table 4.3: Percentage accuracy of prediction for different types of emotion-transition features used for CRF with email dataset before and after using the relative position features

Classifier	k=1	k=2	k=3	k=4	k=5
CRF Type 1	55.07	55.73	55.83	57.5	55.83
CRF Type 2	54.14	55.34	54.86	70.58	55.17
CRF Type 3	52.73	54.09	54.58	69.66	53.64
CRF Type 4	54.64	55.05	55.32	70.08	55.08
CRF Type 5	50.24	51.18	49.39	59.8	50.63
Naïve Baseline	56	56	56	56	56

Table 4.4: Percentage accuracy of prediction for different types of emotion-transition features used for CRF with review dataset before and after using the relative position features

out transition features (Type 1), CRF used with relative position alone outperforms the baseline performance accuracy. Further note that in general there is an increase in prediction accuracy for all the configurations of CRF and for both the datasets. This shows the performance gain from using relative position when used in conjunction with transition features for emotion prediction. Second observation is that Type 3 features have the highest prediction accuracy for email dataset and Type 2 features work the best for reviews. We discuss about this at length in the Section 4.6.

To validate the results, two-tail t-tests were performed and it was noted that the results obtained are statistically significant. The difference between email data set without using position labels and the one using them is significant at level of 0.1557. Further, difference between review data set using CRFs with and without

Transition features used for CRF	Emails Accuracy	Reviews Accuracy
Type 3	48.38	70.58
Type 4, Type 5	43.95	69.94
Type 3, Type 4, Type 5	45.65	69.98
Type 2, Type 3, Type 4, Type 5	42.98	70.48

Table 4.5: Percentage accuracy of prediction for combination of emotion-transition features used for CRF in conjunction with the relative position ($k = 5$ for emails and $k = 4$ for reviews) features.

using position labels is significant at level 0.0035.

Combination of Emotion Transition Features

Further in our study, we evaluated the effect of combining mutliple transition features on the performance of the classifiers. For both the datasets, Type 2, 3, 4, 5 features were together considered for the CRF in addition to the relative position features and the change in predictor accuracy was noted.

The results of combining the transition features for the two datasets are in the Table 4.5. In general, there is no major increase in classifier's performance more than using only Type 2 or Type 3 transition features. This indicates that the performance of CRF with the set of transition features all together is as good as just using the Type 2 or Type 3 transition features.

4.5.2 Accuracy of Standard Classifiers

Next we perform a series of experiments to compare the accuracy of CRF with other standard classifiers. First we consider only transition features without the relative position feature. These results are as shown in Table 4.6. It is evident that Decision Tree and Random Forests perform consistently well across the two data sets. For the reviews dataset, the performance of CRF is comparable to that of Decision Tree.

Classifier	Emails Accuracy	Reviews Accuracy
CRF Type 2	46.37	54.14
CRF Type 3	47.42	53.65
Nave Bayes	61.22	44.19
Decision Tree	61.14	54.98
AdaBoost	45.14	57.1
Random Forests	57.12	52.12
SVM	37.52	49.35

Table 4.6: Percentage accuracy of prediction for different classifiers

Classifier	k=1	k=2	k=3	k=4	k=5
Naïve Bayes	61.22	62	63.05	56.38	59.35
Decision Tree	61.14	62.19	61.23	63.85	58.82
SVM	37.52	63.83	63.16	62.92	61.11
Random Forest	57.12	57.97	59.12	58.92	58.23
Ada Boost	45.14	45.27	44.93	45.37	44.52

Table 4.7: Percentage accuracy of prediction for different types of emotion-transition features used for standard classifiers with email dataset before and after using the relative position features

Even though AdaBoost has very high accuracy for reviews data set, its accuracy is considerably low for emails dataset.

Now we introduce the relative position feature for each of these and observe the change in prediction accuracy. Again, different values of k were used to find out the best way to partition the text so as to achieve maximum prediction accuracy. The results for comparing different values of k are shown in Table 4.7 and Table 4.8.

It is observed that $k = 5$ works the best for emails and $k = 4$ works the best for reviews dataset. Note that introduction of Quarter Labels as a feature for the classifiers tends to improve the prediction accuracy. The improvement is the significant for SVM for email dataset and CRF for reviews dataset.

A major observation from these experiments is that for the reviews dataset, the conditional random fields with Quarter Labels and Type 2 transition features give

Classifier	k=1	k=2	k=3	k=4	k=5
Naïve Bayes	44.2	47.67	47.57	44.2	46.45
Decision Tree	56.38	55.92	54.57	56.39	54.76
SVM	49.29	52.36	51.87	49.29	51.58
Random Forest	52.66	54.11	53.9	52.66	53.45
Ada Boost	57.9	57.2	56.89	57.9	57.8

Table 4.8: Percentage accuracy of prediction for different types of emotion-transition features used for standard classifiers with email dataset before and after using the relative position features

an accuracy of up to 70% which is the highest across all classifiers for all feature sets. The results also reinforce that use of relative position of sentence as a feature improves the performance of classifiers for emotion prediction.

4.6 Discussion

From these experiments we observe that different features work the best for different data sets. Using CRFs for emails dataset, the transition features of Type 3 work the best in conjunction with the use of one-fifth labels ($k = 5$) for the relative position features. On the other hand, for reviews dataset the transition features of Type 2 give highest prediction accuracy when used in conjunction with quarter labels ($k = 4$) for relative position features. We believe that this difference is because Type 3 features capture inclination of users towards certain emotions. When a user writes an email he/she tends to be biased towards the emotions felt at that moment. On the other hand, when writing a review the user is likely to cover all aspects of the product without any inclination towards a particular set of emotions. Another major observation is that CRFs might not perform as well for one dataset as for another. This is discussed at length here.

An interesting insight we gained from our experiments is how the performance of CRF differs for the two datasets. We attribute this to the huge variation of the

Category	Precision	Recall	F1
Happy	0.79	0.32	0.46
Neutral	0.66	0.97	0.79
Tender	0.88	0.34	0.49
Angry	1	0.53	0.7
Sad	1	0.56	0.72

Table 4.9: Precision, Recall and F1 measures for reviews dataset when used with Type 2 transitions and relative position features

sizes of emails with respect to the number of sentences. In particular, out of total 334 emails there are 202 emails with two or less sentences. Together such emails contribute to 21.61% of the dataset. On the other hand the review dataset has only 3 reviews, of total 82 reviews, with 2 or less sentences. In the absence of enough sentences for emails, CRF fails to take advantage of the learnt transition features. Further, reviews by nature tend to be a way of expressing one's emotions.

This can be clearly seen in the following review where a user describes her emotions- *"Natalie Goldberg weaves a wonderful book based on the details of her life and times... I found myself mourning just as she described herself to be at the loss of..."* Here the user is relatively more verbose in expression as compared to many of the emails misclassified by the classifiers, for instance *"Lacy got her eye exam yesterday. the diabetes affected her eyes. I think they might hurt her..."*. The use of an email dataset having customer feedback is more likely to work well for emotion classification using CRFs as it would have users expressing their emotions in a more verbose manner and they are not expected to use very short emails.

Other than prediction accuracy, category-wise precision, recall and F1 measures are also necessary performance measures that give useful insight into a classifier's performance. The observed precision, recall and F1 measures for the five categories of emotions tested on reviews dataset using the best combination of features are shown in the Table 4.9.

The high recall and low precision value for the Neutral category as shown in Table 4.9 means that many of the sentences were misclassified as being Neutral. On the other hand high precision and low recall values of Angry and Sad indicates that many of such sentences were misclassified as another category. This can be fixed by using features that are more specific to Angry and Sad categories. Extending the emotion lexicon is one possible way to do that.

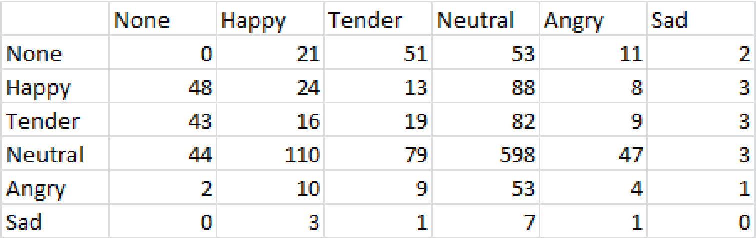
5 Explored Applications

In this section, we discuss a possible application of sentence-level emotion tagging. The experiments performed in this direction are preliminary work towards exploring ways to discover groups of users which can be of particular interest to the company due to their abnormal emotional behavior.

A naïve way to find such a group of users would be to separate users which show abnormally high percentage of say Angry emotion. We believe that finding such a group of users might not be very useful since their angry emotion might be justified in the circumstances faced. However, any person with polite and professional mannerism is likely to end on an agreeable note, with say a Neutral or Tender emotion towards the end of the email. Our study aims to explore if it is possible to find a group of user whose emotional responses deviate from the way majority of users behave.

Tagging emotions for each sentence of a text (say, email) has the advantage that we can study the patterns of changing emotions shown by the user over the short duration of writing an email, from one sentence to another, as well as over a long duration of writing several different emails. In this study, since our focus is on sentence-level emotion tagging, we explore the applications based on short term emotion patterns shown by the user in the duration of an email.

In this context, for every user we use our classifier to classify the emails of several users. For every user, we create an emotion transition matrix. Figure 5.2 shows the emotion transition matrix of one such user. This transition matrix M has six rows and six columns. Five of these stand for the five emotion labels and the sixth one



	None	Happy	Tender	Neutral	Angry	Sad
None	0	21	51	53	11	2
Happy	48	24	13	88	8	3
Tender	43	16	19	82	9	3
Neutral	44	110	79	598	47	3
Angry	2	10	9	53	4	1
Sad	0	3	1	7	1	0

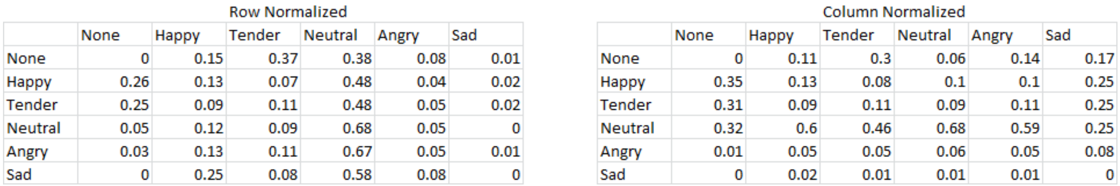
Figure 5.1: Emotion transition matrix of a user

stands for either the start or the end of an email. The emotion of row i is E_i and that of column j is E_j. An entry $M_{i,j}$ of the matrix M stands for the number of transitions from the emotion E_i to the emotion E_j.

Since the number of transitions from one emotion to another can not be used as a standardized metric to compare different users, we normalize the emotion transition matrix. This can be done in two ways. One is to row normalize the matrix where an entry of the row normalized matrix $M^r_{i,j} = M_{i,j} / \sum_{j=1}^{6} M_{i,j}$. This represents the probability of a transition starting from emotion E_i to go to the emotion E_j. Similarly, the other way to normalize the matrix is to use column normalize. In this case, the element of matrix $M^c_{i,j} = M_{i,j} / \sum_{i=1}^{6} M_{i,j}$ represents the probability of a transition ending at the emotion E_j to start from emotion E_i.

We use these two normalized matrices to model the short term emotional behavior of a user while writing the emails. Each of the two normalized matrices contribute for 36 features to compare different users. These 72 features are combined with 5 features corresponding to the percentage of the five emotion tags in the user's emails to constitute a set of 77 features used to model a user's emotional patterns. The next step is to cluster the set of users based on these features to find a set of target group for the company where user shows abnormal behaviour.

A set of 32 users was clustered based on the 77 features using EM algorithm. As a result of this, two clusters were obtained having 10 users for Cluster 1 and 22 users for Cluster 2. The Cluster 1 with less members is the Target group for the



Row Normalized

	None	Happy	Tender	Neutral	Angry	Sad
None	0	0.15	0.37	0.38	0.08	0.01
Happy	0.26	0.13	0.07	0.48	0.04	0.02
Tender	0.25	0.09	0.11	0.48	0.05	0.02
Neutral	0.05	0.12	0.09	0.68	0.05	0
Angry	0.03	0.13	0.11	0.67	0.05	0.01
Sad	0	0.25	0.08	0.58	0.08	0

Column Normalized

	None	Happy	Tender	Neutral	Angry	Sad
None	0	0.11	0.3	0.06	0.14	0.17
Happy	0.35	0.13	0.08	0.1	0.1	0.25
Tender	0.31	0.09	0.11	0.09	0.11	0.25
Neutral	0.32	0.6	0.46	0.68	0.59	0.25
Angry	0.01	0.05	0.05	0.06	0.05	0.08
Sad	0	0.02	0.01	0.01	0.01	0

Figure 5.2: Normalized emotion transition matrices of a user

company since it has less users whose emotional behavior deviates from the normal user behavior. Of the 77 features, 10 features were highly discriminatory when clustering was performed. Some of these features show characteristic properties of the users in the Target group which would help to improve the way company handles human resources and customer feedback. They are as follows-

- Row normalized transition from Happy to Neutral
- Row normalized transition from Happy to End of email
- Row normalized transition from Tender to Neutral
- Row normalized transition from Tender to End of email
- Row normalized transition from Angry to Neutral
- Row normalized transition from Angry to End of email
- Row normalized transition from Sad to Neutral
- Column normalized transition from Neutral to Angry
- Column normalized transition from Start of email to Angry
- Column normalized transition from Neutral to End of email

While different target applications could focus on different emotions and different emotion transition features, for further study we analyze the features that relate to the Angry emotion. This is because the Angry emotion is of particular importance for several scenarios and business models. Also, the fact that four of the ten discriminatory features relate to the Angry emotion indicates that the two sets of users show different characteristics when it comes to the Angry emotion. From further investigation into these results to see behavior of users with Angry emotion we observe the following-

- Row normalized transition from Angry to Neutral is lower for Cluster1
- Row normalized transition from Angry to End of email is higher for Cluster1
- Column normalized transition from Neutral to Angry is lower for Cluster1
- Column normalized transition from Start of email to Angry is higher for Cluster1

This means starting and ending with Angry emotion is more likely for cluster1 (as compared to cluster2), while transition of Neutral emotion to/from and Angry emotion is less likely for this cluster (compared to cluster2). Hence the 10 users of cluster 1 are more aggressive emotionally as compared to the 22 users of cluster 2.

Further investigation into emails of users of the two clusters support these observations. For instance, the following email from a user of Cluster 1 indicates the unexpected change of emotions-

We could take a look at it. I gave it to Mark Taylor and since the fellow lives in England he thought it would be better sent to you. I hate you. That may not be so easy an answer.

Further, our study also shows the interesting result that the naïve way of clustering users based on only the percentage of the 5 emotion categories would not be as useful since the two clusters are observed not to deviate much from each other based on those 5 features. However, for some applications which focus on finding simply the Angry or Sad users instead of users with abnormal emotional behavior, the naïve way might be useful.

6 Conclusion

Emotion classification is a difficult problem, even more when performed at sentence-level. This work introduces two features to model the emotional context of the sentence- transition of emotions in the text and relative position of the sentence in the text. We model the transition of emotions using Conditional Random Fields and evaluate the performance of the new features with several baseline classifiers.

Two major observations can be drawn from this study. First is that different datasets can respond differently to Conditional Random Fields and the proposed transition features depending on characteristics of the data. The second major observation is that relative position of sentence in the text generally improves the emotion prediction accuracy of the classifiers. This means that people tend to express different emotions in different parts of the text.

The features used so far were generic and can be used for any data set having data that contains emotions. In future we plan to extend this work by introducing features that are more specific to the type of dataset being used and hence more specific in application. We also plan to do more fine-grained user profiling based on the transitions of emotions in the text. Such an application would model the emotional behavior of users. This could be of interest to companies that need to analyze user opinion and target user groups for user relationship management.

References

[1] Athena Vakali and Konstantinos Kafetsios. Emotion aware clustering analysis as a tool for web 2.0 communities detection: Implications for curriculum development. In *In Proceedings of the 22nd International World Wide Web Conference*, WWW, 2013.

[2] Chenghua Lin and Yulan He. Joint sentiment/topic model for sentiment analysis. In *InProceedings of the 18th ACM conference on Information and knowledge management*, CIKM'09, pages 375–384. ACM, 2009.

[3] Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *InProceedings of LREC*, LREC, 2010.

[4] P. Koncz and J. Paralic. An approach to feature selection for sentiment analysis. In *InProceedings of the 15th IEEE International Conference on Intelligent Engineering Systems (INES)*, INES'11, pages 357–362. IEEE, 2011.

[5] Y. Zhang, Y. Fang, X. Quan, L. Dai, L. Si, and X. Yuan. Emotion tagging for comments of online news by meta classification with heterogeneous information sources. In *In Proceedings of the 35th international ACM SIGIR conference on research and develpment in information retrieval*, SIGIR, pages 1059–1060. ACM, 2012.

[6] Dipankar Das and Sivaji Bandopadhyay. Word to sentence level emotion tagging for bengali blogs. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, ACL-IJCNLP'09, 2009.

[7] Saima Aman and Stan Szpakowicz. Using roget's thesaurus for fine-grained emotion recognition. In *Proceedings of the Third International Joint Conference on Natural Language Processing*, 2008.

[8] P. Ekman. An argument for basic emotions. In *Cognition and emotions*, pages 169–200, 1992.

[9] Yang Changhua, Kevin Hsin-Yih Lin, and Hsin-Hsi Chen. Emotion classification using web blog corpora. In *Web Intelligence, IEEE/WIC/ACM International conference on IEEE*, WIC. ACM, 2007.

[10] Kunpeng Zhang, Yusheng Xie, Yu Cheng, Daniel Honbo, Downey, Ankit Agrawal, Wei-keng Liao, and Alok Choudhary. Sentiment identificaion by incorporating syntax, semantics and context information. In *Proceedings of the 35th international ACM SIGIR conference on research and development in information retrieval*, SIGIR. ACM, 2012.

[11] Sudheendra Hangal, Monica S. Lam, and Jeffrey Heer. Muse: Reviving memories using email archives. In *Proceedings of the 24th annual ACM symposium on user interface software and technology*. ACM, 2011.

[12] S. Baccianella, A. Esulli, and F. Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of Language Resources and Evaluation*, LREC. European Language Resources Association, 2010.

[13] James W. Pennebaker, Martha E. Francis, and Roger J. Booth. Linguistic inquiry and word count. LIWC. Mahway: Lawrence Erlbaum Associates, 2001.

[14] Narendra Gupta, Mazin Gilbert, and Giuseppe Di Fabbrizio. Emotion detection in email customer care. In *Computational Intelligence*, 2012.

[15] F. Li, C. Han, M. Huang, X. Zhu, Y. J. Xia, S. Zhang, and H Yu. Structure-aware review mining and summarization. In *In Proceedings of the 23rd International Conference on Computational Linguistics*, Association for Computational Linguistics, pages 653–661, 2010.

[16] J. Zhao, Kang Liu, and Gen Wang. Adding redundant features for crfs-based sentence sentiment classification. In *In Proceedings of EMNLP*, EMNLP, 2008.

[17] T. Neylon M. Wells R. McDonald, K. Hannan and J. Reynar. Structured models for fine-to-coarse sentiment analysis. In *In Proceedings of Association for Computational Linguistics*, ACL, 2007.

[18] Saima Aman and Stan Szpakowicz. Identifying expressions of emotion in text. In *Text, speech and dialogue*. Springer Berling/Heidelberg, 2007.

[19] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *InProceedings of ICML*, ICML, pages 282–289, 2001.

[20] Andrew McCallum. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

[21] Enron data set. http://www.cs.cmu.edu/$\sim$enron.

[22] Amazon data set. http://liu.cs.uic.edu/download/data.