Classification of Emotions in Internet Chat: An Application of Machine Learning Using Speech Phonemes

Lars E. Holzman and William M. Pottenger
Computer Science and Engineering, Lehigh University
{leh7, billp}@lehigh.edu
LU-CSE-03-002

ABSTRACT

This article reports our progress in the classification of expressions of emotion in network-based chat conversations. Emotion detection of this nature is currently an active area of research [8] [9]. We detail a linguistic approach to the tagging of chat conversation with appropriate emotion tags. In our approach, textual chat messages are automatically converted into speech and then instance vectors are generated from frequency counts of speech phonemes present in each message. In combination with other statistically derived attributes, the instance vectors are used in various machine-learning frameworks to build classifiers for emotional content. Based on the standard metrics of precision and recall, we report results exceeding 90% accuracy when employing k-nearest-neighbor learning. Our approach has thus shown promise in discriminating emotional from non-emotional content in independent testing.

1. INTRODUCTION

With the information revolution well under way, the degree of communication and number of communication methods is growing rapidly. People converse frequently via a number of mediums. One such medium is Internet chat using various instant messaging clients (e.g., AOL Instant Messenger, MSN Messenger, etc). These communications provide an excellent platform to perform research on informal communications. One such research area that has gained much interest recently is that of tagging the emotion content in informal conversation.

There are several important applications of such tagging. One such application is related to homeland defense. In fact, the research reported herein was motivated by a chat-mining research project we conducted at the behest of the Intelink intelligence network[1]. Intelink is a secure military communications channel used by the US government for critical exchange of information. Intelink's goal is to monitor chat conversation over the network and map relationships between participants and their topics of conversation to determine the appropriateness of usage and the effectiveness of the communication network. They are interested in information such as the frequency of employee communication, the topics discussed, conversational participants and the emotional tone and focus of the conversations. Such information can be modeled using social and semantic networks constructed from chat data [5] [12].

A second important application of emotion tags is the ability to add user feedback to existing instant messaging services. Based on the textual messages produced by a chat participant, an icon or display can represent users' current emotional state as inferred by the system [2].

In addition, the use of emotion detection in user interface design is currently an area of active research. This research is part of the field of Affective Computing first described by Picard [9]. A number of textual, verbal, and non-verbal methods have been applied to allow interfaces and systems to adapt to the emotional state of the user. This area has sparked interest in many fields of research, exciting even those such as Marvin Minsky who has recently written a book on the subject [8].

This paper details a linguistic approach to identifying emotions in chat data. The approach centers on the reproduction of speech from the textual messages logged from an instant messaging client. Speech phonemes present in the message are then modeled to identify the emotion expressed.

This approach offers a number of advantages over more complex methods. The primary advantage is that this method degrades gracefully. Many messages in chat data contain misspellings, do not adhere to grammar rules, and may not contain complete words. We have shown that our approach is robust in the presence of such noise.

The remainder of this paper details our research in the detection of emotion in chat data. In section two, we review related work. In section three, we present our approach. Following this in section four we discuss and provide a detailed analysis of our experimental results. In section five we identify future work, and draw conclusions in section six. We close in section seven with acknowledgements of others who have aided us in this research.

2. RELATED WORK

Much work has been done in the identification of emotion for various applications. The study of emotion for Affective Computing alone is a rather large field. The approaches that exist for lingual emotion detection can be broken into one of a few categories. These categories are: Non-verbal, Semantic, and Symbolic.

2.1 Non-verbal

This category of emotion detection focuses on spoken language. This approach analyzes the characteristics of the speech such as

[1] NSF Grant Number EIA-0070457, Division of Experimental & Integrative Activities.

prosody and spectral information [10]. Using features of spoken language such as the aforementioned ones has proven very successful in identifying emotion in spoken language. This category is the most intuitive. We have all experienced instances where the "tone" of someone's voice indicates to us the emotion they are expressing.

Studies in this category are, in general similar in form to the study reported in this paper. They use a training corpus of sound clips and have domain experts (which in this case are just people fluent in English) declare the emotion class of each sound clip. They then build models out of the non-verbal attributes of the sound clips and test the performance of these models.

While approaches in this category have shown to be very successful they are not easily useful in the domain of chat conversation. There is no non-verbal information associated with the messages that are sent. Furthermore, spoken language is by nature more accurate than typed messages are. Thus, methods to recover from misspoken words are not necessary and would not substantially affect the results.

2.2 Semantic

This category of emotion detection uses an approach that attempts to understand the underlying semantics of language to determine the emotion class. This approach is more complex than the approach described above because it relies upon ability to recognize certain key words in the language. One recent and powerful implementation of this method can be found in Liu, Lieberman, and Selker [6]. This approach uses a large real-world commonsense database to form semantic notions about the story being told in a given sample of natural language. In this case, they integrated the system into an email client to provide feedback to a user.

In the approach affective sentences are extracted from the database. A number of models each representing a different emotion are built from these sentences and the models compete to identify an emotion class for the text. The final models are used to identify the emotion of segments in the text and thus tag them with their proper emotion class.

This method is very robust and attempts to dig deeply into the semantic notion of emotion. This method does rely upon a certain quantity of correct text, though. If the text is malformed or very short it would be difficult to create the models necessary to this approach. Furthermore, this method explicitly chooses to ignore some of the statistical features that show the relationship between emotion and language.

2.3 Symbolic

This category of emotion detection focuses on techniques that attempt to discover patterns in the text that allow emotion tagging. These approaches rely upon syntactical and lexical analyses of the text to discover the emotion class. They attempt to discover clauses and words that identify certain attributes of emotions. They then build models to use these attributes to recover the emotion class.

One such application of this approach can be found in Boucouvalas and Zhe [2]. Their approach utilizes a tagged dictionary to identify the basis of emotion in phrases. It then uses various grammatical features of the phrases to deduce which of the tagged words carries the correct emotion class. As well it resolves syntactical features such as negation and tense. This particular system uses a tag set exceeding the emotion classes. It then uses the attributes to build a scaled model of emotion. Thus, it provides emotion intensity as well as an emotion class.

Another study performs classification in an overlapping domain – emotion act tagging in chat data [12]. This study applies Eric Brill's Transformation-Based Learning (TBL) to the problem of identifying the purpose of messages in chat conversation. Examples of posting acts are: statement, yes-no-question, and emotion. Emotion in this case represents a strong expression of any emotion.

This method uses a set of templates and contextual information to identify emotions. Examples are the use of emoticons[2] and other such expressive measures. TBL uses an iterative error-driven contextual approach to classify the instances using provided templates, which in this case are chat messages. This study addresses the problem of malformed grammar and words through the use of regular expressions.

3. OUR APPROACH

In contrast to the previous methods detailed above our approach employed a method that used very simple and resistant properties of the data. Thus, it is not necessary to make some of the assumptions made by more complex methods such as those of semantic analysis and symbol processing. It attempts to reconstruct the spoken language represented by the chat messages and leverage that information to understand the properties of the language itself. This can be considered closer to a Non-verbal approach.

Furthermore, our approach is local to the message being processed and efficient. A message can be translated into its phonemic equivalent and processed by the model very quickly, which allows real time emotion classification. This is an improvement over models that require intense analysis and model generation. Furthermore, due to the local nature of the method messages can be processed out of order and independent of speaker information or style.

3.1 The Data

The data that we used was obtained from two sources. The first was a set of conversations between one of the authors and another individual over the course of a 2-month period during the summer of 2002. This data set will be referred to as set A. The other source was another set of conversations held by two separate individuals and obtained with the participant's permission for use in these experiments. This set will be referred to as set B and was substantially smaller than set A. It was obtained primarily for testing the extendibility of a model trained only on set A.

These conversations were broken up by message. A message was defined as all the text that was entered before a send message

[2] An emoticon is used to explicitly express an emotion using ASCII characters representing a facial expression and is common in chat conversation.

procedure was invoked (as by pressing a "send message" button or the enter key). Thus, it may take one or more messages to form a semantically coherent statement (this occurs rather frequently, in fact). Manual cleaning was performed and messages that contained no characters or only punctuation were removed. The other messages were manually tagged with an emotion class. The cleaned tagged data was stored in a truth file using the Truth-File Format described in [4]. After this cleaning set A contained 1016 messages and set B contained 185 messages.

3.2 The Training Set

Each message was analyzed separately to discover its phonetic properties. The American English standard phoneme set contains 49 phonemes, which represent the major units of sound used in American English speech including accents and pauses. The Microsoft Speech SDK version 2 (beta) [7] was used to "speak" the message. The phonemes were then extracted from this speech reproduction of the text. This allowed tallies to be generated of each phoneme. These phoneme tallies provided a characterization of the language represented in the textual message. This process was performed by the TDMAPI infrastructure [4]. This allowed processing of the truth-files and acted as a driver for the speech API.

Statistical measures characterizing the messages were used in some of the experiments. These measures allowed further characterization of the language in the message. The measures used were the following: word count, length of the longest word, length of the shortest word, average word length, total message length, number of periods present, number of exclamation points present, and the number of question marks present. In these measures words were defined as units of text separated by white space or non-alpha numeric characters (i.e. punctuation). This is a simplified version of a "graphical word". These statistical measures were considered useful and relevant because they describe the linguistic features of the message.

A number of different classification methods were used to characterize the emotion content of the messages. They can all be derived from one original set of emotions. The original set of emotions is one common to the recognition of emotion from facial expressions. The emotion set consists of: Neutral, Angry, Sad, Afraid, Disgusted, Ironic, Happy, and Surprise. This was the same set used by Polzin [10].

Neutral is used in the place that there is no emotion present in the message or that there is no emotion discernable in the message. Emotions are in general not defined by a set of conditions but instead occur on a continuous and overlapping scale and thus there are many cases where there is no clear emotion. Only messages with one clear emotion were tagged as such. Thus, the neutral classification as well contains the indeterminate category.

It was discovered that these classes originally derived for facial expressions did not represent that chat conversations being analyzed well. As such, the class distribution was very imbalanced and in fact some tags never occurred or occurred too infrequently to be used in the study. Table 1 contains the distribution of classes in all the training data.

To accommodate this distribution for some of the experiments a different class configuration was used. This configuration divided the data into two basic classes: neutral and emotional. The neutral class was the same as for tagging, the emotional class contained all other classes. This created a more reasonable distribution with 942 (78.43%) neutral entries and 259 (21.57%) emotional entries. This classification method is referred to as binary emotional classification.

Table 1. Frequency of classes in the data

Class	*Frequency*	*% of Total*
Neutral	942	78.43%
Angry	59	4.91%
Sad	14	1.17%
Afraid	0	0.00%
Disgusted	11	0.92%
Ironic	23	1.92%
Happy	124	10.32%
Surprise	28	2.33%
Total	1201	

As well, for some experiments all the neutral entries were removed from the dataset. This allowed the ability to learn the actual emotion classes without the noise and imbalance introduced by the highly represented neutral class. The final set of attributes present in the training set is detailed in Table 2.

Table 2. Attributes present in the training set

Attribute	Type
Phoneme 1 (-) count	Numeric
Phoneme 2 (!) count	Numeric
…	…
Phoneme 10 (aa) count	Numeric
Phoneme 11 (ae) count	Numeric
…	…
Phoneme 49 (zh) count	Numeric
Minimum Word Length	Numeric
Maximum Word Length	Numeric
Average Word Length	Numeric
"." occurrence	Numeric
"!" occurrence	Numeric
"?" occurrence	Numeric
Total Word Length	Numeric
Word Count	Numeric
Emotion Class	Nominal

3.3 Attribute Selection

After a suitable training set was built the connection of the phonetic information and emotion classes was established through the use of statistical measures and machine learning analysis. This process was used to verify the validity of this approach independent from the ability to use this method for chat emotion detection. All attribute selection was performed on the merged training set containing the data from both conversation sets.

The first attribute selection method that was employed was the simple application of a Chi-Squared statistical metric to determine the dependence of emotion class on each of the attributes. Attributes with a chi-squared value greater than zero were considered as related to emotion classes. At this stage we were not looking for strong connection because this was only a preliminary screen.

Experimentation was conducted by performing a number of attribute selection methods implemented in the WEKA [11] library on the training set. Both wrapped and unwrapped methods were used. Included in the unwrapped methods was the chi-square metric, info gain metric, and subset evaluation. The wrapped methods used decision trees and Bayesian approaches to evaluate attribute sets. The space was searched using various optimization methods. This allowed us to find a set of attributes that many machine learning approaches found useful for building a model. The methods were allowed to vote on attributes and only attributes that appeared in the majority of methods were considered useful.

3.4 Machine Learning

Finally, machine learning was applied to build a model suitable for the identification of emotion classes using the attributes provided. A number of attribute sets was used to discover the optimal approach to using phonemes to build a model of emotions. The sets used were the full attribute set as described above, an attribute set consisting only of the phonemes, and the attribute set of highly correlated attributes as described in the previous section.

Three sets of experiments were performed in the machine learning stage. The first experiment set that was performed was the application of machine learning to the data set using the full range of emotional classes. These experiments tested the ability of machine learning to learn the concept of specific emotion classes (specifically Happy and Neutral which were the only classes well represented). The second experiment set that was performed involved the application of machine learning to the data set using the binary classification method. This experiment tested the ability of machine learning to learn the concept of an emotion expression independent of the emotion type. The third experiment that was performed using the data set that had all the Neutral entries removed. This experiment tested the ability of machine learning to differentiate between emotion classes within a set of affective entries. This experiment was only performed on the full attribute set.

The primary machine learning method used for this application was k-nearest neighbor instance based learning (IBk) [1]. This method is very simple and relies upon using a metric to find k instances close to the instance being studied. The instance is classified by majority vote of the k nearest neighbors. These instances can be weighted and in this study were weighted by the inverse of their attribute distance from the unclassified instance. This weighting allows the algorithm to modify the strength of the votes by their distance from the unclassified instance. This method was studied because it has the ability to model very complex domains and scales well. The only limit to the complexity of this machine learning method is the limit on the ability to store instances. As well, for the purposes of this application it was not necessary to discover the exact manner in which classification occurs.
The number of nearest neighbors necessary for correct classification was unknown. A study of this parameter was conducted to find the optimal value for the classification experiments performed in this domain.

3.5 Evaluation

We used the standard evaluation metrics of Precision (*TP/TP+FP*) and Recall (*TP/TP+FN*) where TP is the number of true positives, FP is the number of false positives, and FN is the number of false negatives. As well, the metric of F_β which represents the harmonic mean of precision and recall was used. In all cases $\beta=1$ which gives equal importance to precision and recall. It is not clear which is more important in this application so a balance was striven for. This is reflected in higher overall F_β scores.

Ten-fold cross validation was used for all testing on the combined data set and the cross-emotional test. As well, cross testing (with set A for training and set B for testing) was performed to allow testing on novel data with different speakers.

4. RESULTS

In this section the result that there is a connection between phoneme counts and emotion class is presented. As well, it is shown that this correlation can be used successfully to identify emotional content and discriminate between emotions to some degree.

4.1 Statistical Correlation

The chi-squared metric identified a number of attributes including many phoneme counts as being related to the emotional classes. The results of this analysis can be found in Table 3. Only results with correlation greater than 50 are provided. 14 attributes which had a chi-squared metric of greater than zero (of which 13 are phoneme counts) are omitted from the table. These were still considered related because they showed correlation between a phoneme count and the emotion class.

Table 3. Chi-squared results > 50

Attribute	χ^2
Phone. 26 (h)	213.98
Total Length	139.81
Word Count	135.57
Max Word Len.	115.99
Phone. 7 (_)	86.01
Phone. 41 (t)	80.74
Min Word Len.	76.62
Phone. 33 (n)	65.50
Phone. 27 (ih)	65.22
Phone. 12 (ah)	56.65

These results suggest that there is in fact a relation between the phoneme counts and the emotion class. The attribute most correlated with the class was a phoneme count. Furthermore, nineteen distinct phoneme counts showed a relation to the

emotion class. The other attributes that were selected model features of the language such as the length of words.

4.2 Attribute Selection

Attributes were selected by a majority vote of automated attribute selection methods. Seven methods were used and an attribute had to be selected by at least four of the methods to occur in the final subset. Using this method the attributes selected were: phonemes 12, 16, 21, 23, 26, 27, 30, 33, 37, total message length, minimum and maximum word length, and question mark occurrence. This result again verifies that there is a substantial connection between the phoneme counts and the emotion class.

4.3 Machine Learning

It was verified that k-nearest-neighbor instance based learning performs better in this domain than other commonly used methods such a decision trees. Therefore, only IBk was explored in depth. Figure 1 displays the evaluation of a number of machine learning methods (Naïve Bayes, One R, Decision Table, Ib1, j48, Ib20) as implemented in the WEKA library [11]. These methods were selected because they represent a survey of commonly used off-the-shelf machine learning methods. The methods were evaluated on their F_β scores using ten-fold cross validation.

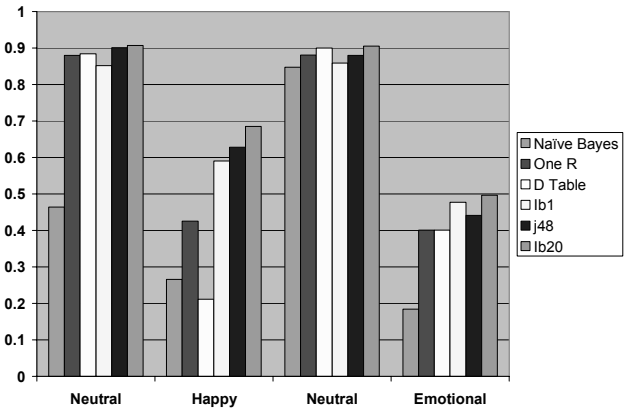

Figure 1. A Survey of Machine Learning Methods

It is important to note that k=20 is not the optimized value of k. The figure illustrates clearly that IBk (without an optimal value of k) performs at least as well as the surveyed machine learning methods. This paired with the advantages of IBk detailed above suggest IBk is the ideal machine-learning method for this application.

4.4 Full training set

The following results were created from the use of machine learning on the entire training set containing the instances from both set A and set B. Again, stratified ten-fold cross validation was used to generate all results reported.

4.4.1 Neutral vs. Happy

In this experiment the ability of machine learning to tag data with the emotion classes happy and neutral (which were the two largest classes by a large margin) was assessed. It was shown that IBk performed reasonably well for this application. These experiments were conducted on the tagged data with all emotion classes represented. Classification of the other classes was however not studied.

To optimize IBk for this application k was varied over a range of values from ten to forty-two. These values were obtained by first setting a floor of ten (because of the complexity of this domain anything less would have been insufficient). Next, k was incremented until a peak or stability point was found. This point was found to be less than or equal to forty-two for the experiments performed. Figures 2 and 3 show example precision, recall, F_β graphs used to optimize k for this experiment with the full attribute set. Table 4 presents the results of this experiment for each of the attribute sets. The optimal value of k is considered the lowest value at which the results obtain the maximum F_β. The only exception to this rule is if there is some k which represents an overlap of the maximum values for both classes (i.e., it may not be the first maximum for one of the classes). The charts provided are for the full attribute set but are indicative of other attribute sets.

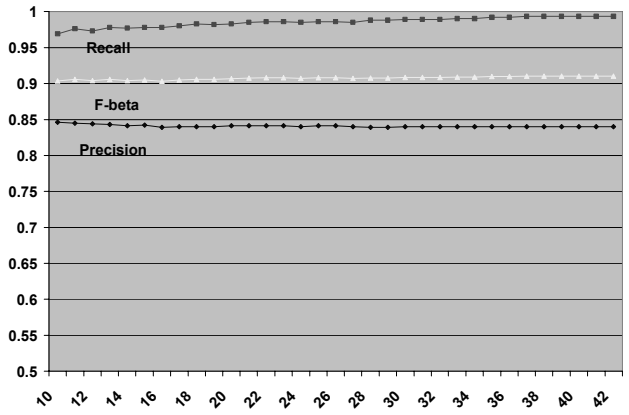

Figure 2. Neutral performance vs. k in IBk

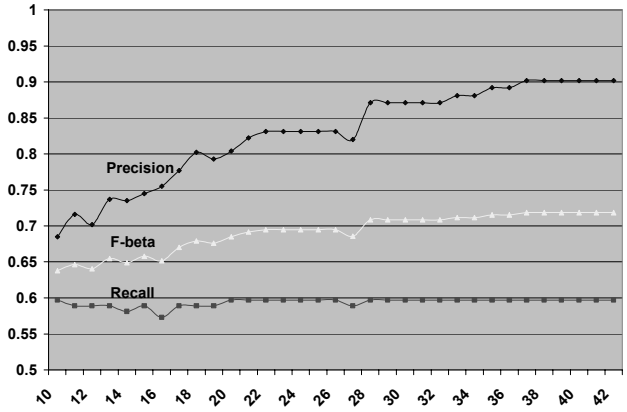

Figure 3. Happy performance vs. k in IBk

Table 4. Neutral vs. Happy results

		Neutral			*Happy*			
		Prec.	**Recall**	**F-beta**	**Prec.**	**Recall**	**F-beta**	**k**
Full	**Avg**	0.841	0.985	0.9074	0.829	0.594	0.6911	
	Max	0.840	0.993	0.9101	0.902	0.597	0.7185	37
Phonemes	**Avg**	0.840	0.978	0.9035	0.778	0.606	0.6814	
	Max	0.840	0.979	0.9042	0.792	0.613	0.6911	23
Subset	**Avg**	0.839	0.987	0.9072	0.859	0.595	0.7030	
	Max	0.841	0.993	0.9101	0.902	0.597	0.7185	37

There are some significant findings to be noted from Table 4. The first is that F_β for Neutral has a maximum of .902 and the F_β for Happy has a maximum of .72. These are significant results. Furthermore, the maximum values of F_β for the Neutral and Happy class occur at the same k value. This implies there is an optimum k for classification purposes. The best results occurred for the full attribute set and selected attribute set.

4.4.2 Neutral vs. Emotional

The results of this experiment showed that the machine learning using the binary classification method was less stable. A plot of the performance variation as k is varied can be found for the happy class in Figure 4. Neutral behavior for the full attribute set was similar to the behavior exhibited in chart 2 showing a rather stable reaction to the variation of k. The results for each of the attribute sets are summarized in Table 5.

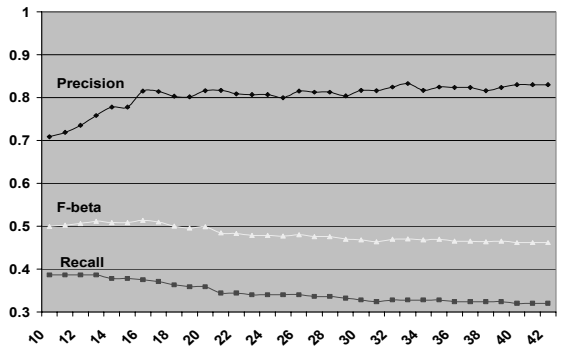

Figure 4. Emotional performance vs. k in IBk

Table 5. Neutral vs. Emotional results

		Neutral			*Emotional*			
		Prec.	**Recall**	**F-beta**	**Prec.**	**Recall**	**F-beta**	**k**
Full	**Avg**	0.844	0.977	0.9056	0.804	0.345	0.4822	
	Max	0.851	0.977	0.9091	0.815	0.375	0.5137	16
Phonemes	**Avg**	0.843	0.976	0.9044	0.793	0.338	0.4739	
	Max	0.844	0.980	0.9064	0.802	0.344	0.4815	31/18
Subset	**Avg**	0.840	0.980	0.9049	0.825	0.323	0.4632	
	Max	0.842	0.984	0.9075	0.850	0.382	0.4733	23

The performance of the Emotional class studied in these experiments was not as significant as that of the Happy class in the previous experiments. Nonetheless, it was substantially greater than chance and thus a significant result nonetheless. The lower result can be attributed to the fact that many of the less represented emotions are part of the emotional class. Thus, since these emotions are not well represented they could not be learned. This assumes that IBk is still looking for particular emotions in the emotional classification, though. Precision remained high in these experiments. The best results occurred for the full phoneme set.

4.5 Cross-testing

This approach used the instances from set A for training and the instances from set B for testing. This experiment was used to test if a model learned on one set of conversations would scale well to other conversation with unique participants and styles.

4.5.1 Neutral vs. Happy

This experiment was the parallel of the experiment described in section 4.4.1. In this experiment, though, instances from set A were used to classify the instances from set B. Due to the relatively smaller size of set B the result show greater variation. This is simply because the change of one TP can affect precision and recall more with less overall instances. The performance plot for Happy as k is varied can be found in Figure 5. A summary of the results can be found in Table 6.

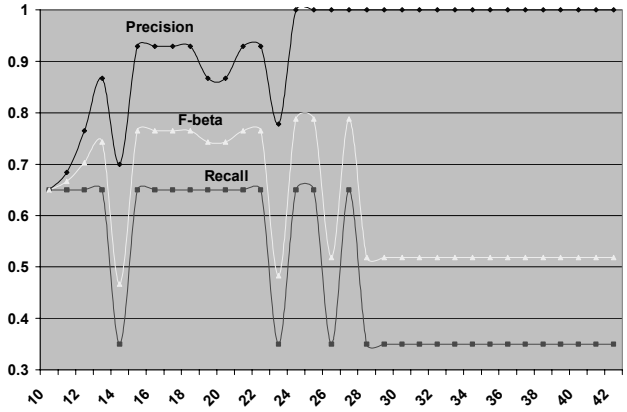

Figure 5. Happy performance vs. k in IBk

Table 6. Neutral vs. Happy results

		Neutral			*Happy*			
		Prec.	**Recall**	**F-beta**	**Prec.**	**Recall**	**F-beta**	**k**
Full	**Avg**	0.840	0.994	0.9109	0.932	0.486	0.6196	
	Max	0.855	1.000	0.9218	1.000	0.650	0.7879	27
Phonemes	**Avg**	0.861	0.988	0.9202	0.911	0.384	0.5380	
	Max	0.865	1.000	0.9276	1.000	0.395	0.5663	30
Subset	**Avg**	0.828	0.984	0.8992	0.781	0.206	0.3233	
	Max	0.851	0.973	0.9079	0.765	0.342	0.4727	10

These results were very significant. In many cases the precision or recall approached or actually reached 1. This is an ideal result. The results for this cross-test experiment actually exceeded the results of the cross-validated experiment. This shows that the model learned is scalable for learning Neutral and Happy.

4.5.2 Neutral vs. Emotional

This experiment was the parallel of the experiment described in section 4.4.2. This experiment, as well, used instances from set A to classify the instances from set B. The performance plot for Emotional as k is varied can be found in Figure 6. A summary of the results can be found in Table 7. It is interesting to note that although the precision is high (exceeding .9) for Emotional in this application, the recall trade off is too great and thus since we are optimizing F_β the results reported are not optimal for either precision or recall. This is true in many of the results reported for this experiment (i.e., that precision and recall can be optimized locally but are significantly lower at the F_β max).

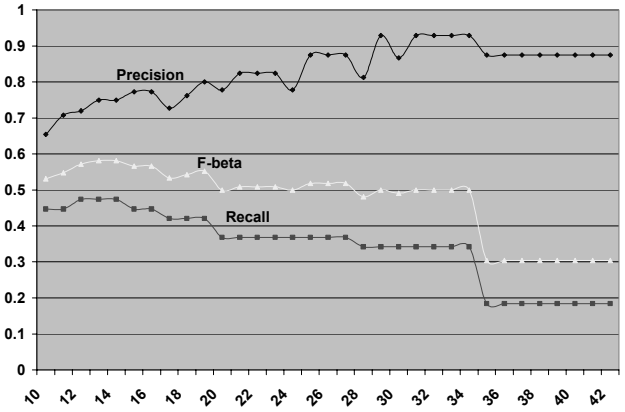

Figure 6. Emotional Performance vs. k in IBk

Table 7. Neutral vs. Emotional results

		Neutral			*Emotional*			
		Prec.	**Recall**	**F-beta**	**Prec.**	**Recall**	**F-beta**	**k**
Full	**Avg**	0.853	0.979	0.9112	0.830	0.342	0.4714	
	Max	0.854	0.993	0.9183	0.750	0.474	0.5809	31/13
Phonemes	**Avg**	0.843	0.976	0.9044	0.793	0.338	0.4739	
	Max	0.844	0.980	0.9064	0.802	0.344	0.4815	31/18
Subset	**Avg**	0.840	0.980	0.9049	0.825	0.323	0.4632	
	Max	0.842	0.984	0.9075	0.850	0.328	0.4733	23

These results did not exhibit the coherence of the other results reported. As is exhibited in Figure 6 the maximums did not occur at the same point (although local maximums could be found at both points in most cases). This is probably due to the inability to develop adequate models for the less represented emotions because of their scarcity in the data. Nonetheless, performance peaked at over 90% F_β for the Neutral class, and at over 80% average for precision of the Emotional class.

4.6 Emotion-only testing

This experiment attempted to determine the ability of IBk to differentiate between the emotion classes with the majority class of Neutral no longer present in the data set. Thus, it examines the differentiation of phoneme representation in emotion expressions. The classes for this experiment were thus: Angry (22.8%), Sad (5.4%), Disgusted (4.25%), Ironic (8.88%), Happy (47.9%), and Surprise (10.8%). Results were only reported for the well represented classes of Angry, Happy, and Surprise. The training set contained all classes, though. This experiment used only the full attribute set. Ten-fold cross validation was used to obtain the results reported. Figure 7 plots the F_β performance of each of the classes as k is varied. Figure 8 provides a result summary. It can be seen that when k is around thirty-one there is a peak of the F_β ratings. While these F_β ratings are not individually optimal for the measure it seems clear that this is the globally optimal value for k.

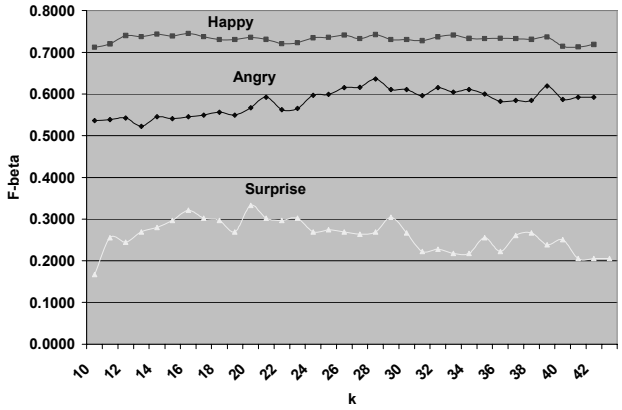

Figure 7. Emotions F_β vs. k in IBk

Table 8. Emotion distinction results

		Prec.	**Recall**	**F-beta**	**k**
Happy	**Avg**	0.528	0.647	0.581	
	Max	0.575	0.712	0.636	28
Angry	**Avg**	0.643	0.850	0.732	
	Max	0.665	0.847	0.745	16
Surprise	**Avg**	0.321	0.230	0.263	
	Max	0.346	0.321	0.333	19

These results reflect that IBk is successful at differentiation emotions (given proper representation in the training set) at a rate well above chance. Therefore, IBk is both successful at extracting major emotion classes from background neutral entries and at differentiating emotion entries without neutral entries present. The algorithms success is still restricted to well represented emotion classes, though.

5. FUTURE WORK

These results were successful in showing that in principal phonemes can be used to classify emotions in chat conversation. The next stage of research concerns extending this result. Further training set development must occur to learn a more general model of language. This development would require tagging of a more diverse set of chat documents including other informal communication domains such as IRC and e-mail. This will allow the emotion classes not well represented in the training set used for this study to be further studied.

Furthermore, further tests will be conducted to determine the extensibility of such a system. Tests of this nature will include test such as romantic language testing, testing on domains other than chat, and cross testing models between domains and language. Romantic language testing will attempt to extend these results to languages with similar phoneme sets to English (for instance, Spanish or French). Testing as well will be conducted on other dialogue driven domains such as drama and literature dialogue. Finally, models from one domain will be tested on another to show the extensibility of a model built using this method.

The eventual goal of this research is to incorporate a phoneme-driven emotion detection system into a more complex architecture. This architecture could be used to build complex models of chat conversation and allow access to information such as the thread structure, the participant interactions, etc.

6. CONCLUSIONS

The experiments provided in this paper show the potential of reconstructing speech from Internet chat dialogue and using attributes of that speech to detect emotion. K nearest-neighbor Instance Based Learning was found to be the ideal method to learn this concept. The IBk machine learning method used a training set that consisted of instances containing phoneme counts for each of the American-English Phonemes as well as other statistical measures. Training sets consisting of only phoneme counts proved useful, although they did not perform as well as instances containing other statistics, as well.

It was also shown that the number of nearest neighbors (k) can be optimized to improve performance. The optimal value of k was not found to be constant but was almost always in the range of twenty to forty. The optimization was performed using F_β as a metric to allow optimization of both precision and recall. As well, it was shown that use of the complete attribute set (in contrast to an automatically selected attribute subset) was optimal for this application.

7. ACKNOWLEDGEMENTS

The authors gratefully acknowledge the help of family members and friends. Co-author Lars E. Holzman would like to acknowledge and thank co-author William M. Pottenger and the members of the Distributed Textual Data Mining lab for their support in accomplishing this work. Co-author William M. Pottenger gratefully acknowledges his Lord and Savior, Yeshua the Messiah (Jesus Christ), for providing him with salvation.

8. REFERENCES

[1] Aha, D., and D. Kibler. Instance-based learning algorithms. *Machine Learning*, vol.6, pp. 37-66. 1991

[2] Boucouvalas, A. C., Zhe, Xu. Real Time Text-To-Emotion Engine for Expressive Internet Communications. *International Symposium on CSNDDSP*. 2002.

[3] Devillers, Laurence, Vasilescu, I., Lamel, L. Annotation and Detection of Emotion in a Task-oriented Human-Human Dialog Corpus. *Proceedings of the Workshop on Dialogue Tagging for Multi-Modal Human Computer Interaction.* 2002. Can be found on the WWW: (http://www.research.att.com/~walker/isle-dtag-wrk/*)*

[4] Holzman, L.E. Fisher, T.A., Galitsky, L.M., Kontostathis, A. and Pottenger, W.M. A Software Infrastructure for Research in Textual Data Mining. *Currently under review for ACL 2003*. 2003.

[5] Khan, F.M., Fisher, T.A., Shuler, L., Wu, T., Pottenger, W.M. Mining Chat-room Conversations for Social and Semantic Interactions. Can be found on the WWW: http://www.cse.lehigh.edu/techreports/2002/LU-CSE-02-011.pdf. 2002

[6] Liu, H., Lieberman, H., Selker, T. A Model of Textual Affect Sensing using Real-World Knowledge. *Proceedings of the Seventh International Conference on Intelligent User Interfaces (IUI 2003),*pp. 125-132. 2003.

[7] Microsoft Speech SDK. Can be found on the web at: http://www.microsoft.com/speech.

[8] Minsky, M. The Emotion Machine. *Not yet available in hard copy.* http://web.media.mit.edu/~minsky. 2003

[9] Picard, Rosalind W. Affective Computing. Media Laboratory, Perceptual Computing TR 321, MIT Media Lab. 1995.

[10] Polzin, Thomas S., Waibel, A. Emotion-Sensitive Human-Computer Interfaces. *Proceedings of the ISCA Workshop on Speech and Emotion.* 2000.

[11] Witten, I.H., Frank, E. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, San Francisco, CA. Available on WWW: www.cs.waikato.ac.nz/ml/weka

[12] Wu, T., Khan, F.M., Fisher, T.A. Shuler, L.A. and Pottenger, W.M. Posting Act Tagging Using Transformation-Based Learning. In the *Proceedings of the Workshop on Foundations of Data Mining and Discovery*, IEEE International Conference on Data Mining (ICDM'02). December 2002.