A Hybrid AI Approach to the Classification of Emotive Text

Curry I. Guinn, *University of North Carolina Wilmington*

***Abstract*—In this paper we will discuss Natural Language Processing (NLP) techniques used to classify the emotional content of a segment of autobiographical text. In particular we will examine 1) statistical training techniques which rely on Bayesian models for classification, and 2) statistical analyses which rely on hand-built annotations of emotive and content keywords. In isolation, each methodology performs to a relatively high level of precision (85-90%). This paper then looks at how to combine these two approaches to produce an overall system which has an accuracy that approaches 98% for a corpus of over 285 autobiographical entries.**

***Index Terms*—Natural language processing, affective computing, statistical NLP, Linguistic Inquiry and Word Count (LIWC) text analysis software.**

I. Introduction

Affective computing has emerged as an important area within computer science as more user-centric applications are developed [1]. Many modalities are being explored to ascertain the emotional state of users: facial gestures, spoken vocal affect, skin conductance, heart rate, respiration, posture, hand gestures, body movements, and other physiological measures. This paper describes research in processing the word content of a user's dialogue, whether spoken or written. Our intent is to incorporate this work with previous research done in our lab on affect recognition using facial gesture by taking audio-visual transcription of a subject while they are talking [2]. Our task for the research project described here is to determine whether a relatively short span of text is indicative of a positive affect or a negative affect by the speaker (in our study, the mean number of words in each sample is 102; for more on the domain of study, see Section II). The author has done prior research in affect recognition using text where the primary mechanism involves using semantic grammars which have been tagged with emotional content [3]. The process of hand-tagging grammars with emotional data is time-consuming and an ultimately brittle method. While this methodology can be suitable for relatively small domains, it is not suitable to open-ended domains with a vast breadth of vocabulary and grammar use. This study will investigate the use of statistical methods of classifying text as well as the use of large hand-built dictionaries that have been tagged with emotional (and other) data.

In this paper we will show how corpus data was collected for our study in Section II. In Section III, we will explore the use of probabilistic learning algorithms to classify texts in the domain based on Bayesian models. In Section IV, the use of an extensive hand-built dictionary used by text analysis software on this domain will be explored. In Section V, these two techniques will be combined to produce a system that functions better than using either technique alone. In Section VI, we will discuss areas for further research.

Manuscript received March 11, 2009.

Curry I. Guinn is with the Department of Computer Science, University of North Carolina Wilmington, Wilmington, NC, USA, 28403. Phone: (910)-962-7937, Fax: (910)-962-7457, email: guinnc@uncw.edu.

II. Domain Collection

Our goal is to determine whether a speaker/author is displaying either positive or negative affect in a relatively small sample of text. The data that we used for this study was gathered in an independent study conducted prior to this investigation. This corpus comes from a study done by Lecci and Wirth to determine the effect on music on the retention of emotional state [4]. In the process of conducting the study, Lecci and Wirth wished to induce a positive or negative mood in the subject. They used a well established mood induction technique where subjects are asked to write about an event in their lives that made them either *very sad* or *very happy* [5]. A portion of the instruction to subjects is given in Figure 1.

Figure 1 Instructions to Subjects in Lecci/Wirth Study

This study is intended to research how music affects our ability to recall positive and negative memories. Therefore, we would now like you to think about some event in your life that was very sad. This need not be a recent event, but we would like you to indicate approximately when this event occurred. Sometimes other feelings occur along with sadness, but we would like to know something that happened to you that made you feel mostly sad. Take a few moments to try and think of such an event in your life.

Now we would like you to write about this event in as much detail as you can remember. We would like you to emphasize the exact thoughts and feelings you experienced and the negative things that happened to you because of this sad event.

Subjects were given about 3 minutes to hand-write a response. The subject pool consisted of evenly divided male and female undergraduates at a state university. Overall, 145 NEGATIVE and 143 POSITIVE autobiographical entries were recorded. Sample entries are given in Table 1.

Table 1 Examples of Positive and Negative Autobiographical Entries in the Lecci/Wirth Study

Positive	Negative
About 4 months ago, I achieved one of my biggest and hardest worked at goals -- I graduated as valedictorian of my senior class with a 4.0 gpa. I remember feeling excited and a rush of energy as graduation drew near and I realized my personal victory. What was best about it was the fact that it was personal, not a goal that my parents or anyone else pushed for or expected. I did it on my own. That was its own reward.	*The event that caused me to be the saddest in my life was when my parents separated. I was approximately 14 yrs. Old and it was my freshman year in high school. My parents were have*[sic] *marital problems due my father's infidelity and financial strains. I was forced to be a mediator for my parents and be strong for my two younger male siblings. I had no one to confide in and convey how I was feeling. I was very sad, lonely, and* [time limit exceeded]
I can recall in 1985 when my grandparents took my sister and I to the Rocky Mountains. We went to Pigeon Forge and drove to the top of a mountain where we had a picnic lunch. It was fall and all the leaves were bright colors: red, orange, yellow. My grandfather gathered a big pile of leaves together and began to jump around in them. He buried my sister and I in them.	*One sad event that took place in my life was when my best friend of fourteen years betrayed me. I had a crush on this guy who ended up asking my friend out. She knew all about my crush and knew it would hurt my feelings so she asked my opinion. I told her I couldn't make that decision for her. So she went out with him anyway. I felt ugly, angry and totally* [sic]. *I cried for weeks over this event but not because of the guy but because I didn't feel like I could trust my friend anymore. We still aren't friends like we use[time limit exceeded]*

For the purposes of our study, 3 NEGATIVE responses were thrown out because the subjects did not follow the correct protocol and wrote about something other than what was asked. 285 responses are in the full data set used for the analysis. The median number of words per subject entry is 102.5; the mean is 101.6 with a standard deviation of 33.1. A summary of texts broken down by POSITIVE and NEGATIVE categories is summarized in Table 2.

Table 2 Statistics on the Autobiographical Text Entries Broken Down by Affect Category

Affect Category	Median	Mean	Standard Deviation
NEGATIVE (n=142)	108	105.95	33.56
POSITIVE (n = 143)	94	97.11	32.21
Overall (n = 285)	102.5	101.6	33.14

III. N-GRAM ANALYSIS

Statistical analysis using Bayesian statistics is a standard technique for a number of classification tasks within natural language processing [6]. This author has made extensive use of Bayesian statistics for categorizing the semantic class of utterances [7]. To use this technique for single words, a corpus is divided into a training set and a test set. Using the training set, conditional probabilities are generated for each word and each category. For our domain, we are asking, given a particular word what is the probability that it occurs in a NEGATIVE text and what is the probability that it occurs in a POSITIVE text? Then, for each instance in the test set, we use those probabilities to predict the category of the instance based on the words the test instance contains.

Because the subjects in the Lecci/Wirth study were primed in their instructions with the words “very sad”, “sad”, “very happy”, and “happy”, we deleted these words (and words that can be derived from them like “happier”, “saddest”, etc.) from our training and test data. A number of subjects began their entries with something like “the event that made me the saddest”, etc. This deletion, of course, makes the categorization task more difficult, but their deletion removes the bias introduced by the instructions given to the subjects.

A. Single Word Analysis

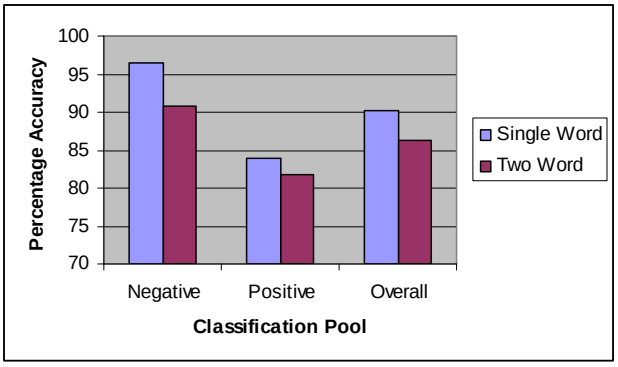
Single word analysis involves looking at the conditional probability of each word and each category. We employed a leave-one-out training methodology where the statistics are learned by analyzing “all but one” instances of the data, and then seeing if the system can correctly classify the unseen instance. Then those probabilities are thrown away, and the system relearns the probabilities on “all but a different one” and the system tries to classify that different unseen instance. This is repeated throughout all the data. Using a leave-one-out training methodology, the system was able to correctly label NEGATIVE responses as NEGATIVE

96.48% of the time and POSITIVE responses 83.9% of the time. The total accuracy was 90.754%

B. *Two Word Analysis*

Two word analysis involves looking at the conditional probability of each **adjacent pair** of words and each category. Given the relatively small data set, it was not clear that there was sufficient data to use adjacent word pairs. The analysis shows that word pairs are still effective, but not quite as effective as single words for this training set. Using leave-one-out training, word pairs were able to correctly label NEGATIVE responses 90.8% of the time; POSITIVE responses, 81.8%; for an overall score of 86.3%. A summary of the Single Word and Two Word analyses is given in Figure 2.

Figure 2 The Percentage of Correct Classifications using Single Word and Two Word on Negative and Positive Texts



Bayesian analysis using single words was an effective tool for predicting whether the unseen text was in the NEGATIVE category. However, it gave a substantial number of errors when analyzing POSITIVE texts (rating them as NEGATIVE). Human analysis of those falsely labeled POSITIVE texts showed that the content would easily be judged as emotionally positive by a human judge. The statistically trained system simply lacked an adequate training corpus to handle the broad expression of positive emotions and positive events. (Another plausible explanation for the better performance for NEGATIVE texts is that the average length of a NEGATIVE text entry is 8-14 words longer than POSITIVE text entries. The more data, the better the statistics. However, the substantial increase in performance for NEGATIVE entries does not seem likely to be caused by this modest increase in size.)

IV. HAND-BUILT LINGUISTIC ANALYSIS TOOLS

One solution to the lack of an adequate training corpus is to use a dictionary of words that have been pre-labeled with Positive, Negative, and other categories. Such a dictionary does exist, the Linguistic Inquiry and Word Count text analysis software [8]. The dictionary used by LIWC has been developed and refined over two decades [9]. LIWC produces a substantial amount of analysis:

> *This data record includes the file name, 4 general descriptor categories (total word count, words per sentence, percentage of words captured by the dictionary, and percent of words longer than six letters), 22 standard linguistic dimensions (e.g., percentage of words in the text that are pronouns, articles, auxiliary verbs, etc.), 32 word categories tapping psychological constructs (e.g., affect, cognition, biological processes), 7 personal concern categories (e.g., work, home, leisure activities), 3 paralinguistic dimensions (assents, fillers, nonfluencies), and 12 punctuation categories (periods, commas, etc)*[9].

The LIWC2007 dictionary has over 4500 words and word stems.

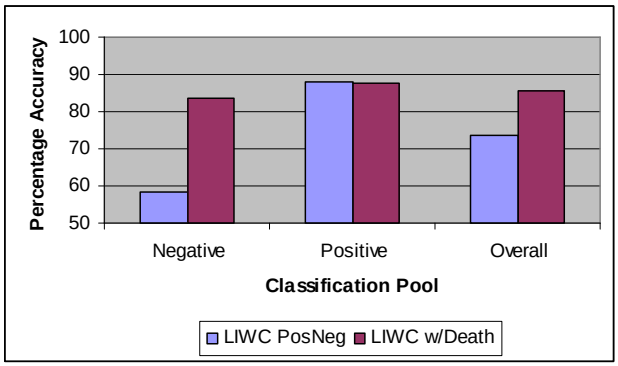
A. *LIWC's POSEMO and NEGEMO Categories*

Two of LIWC's categories are of particular interest to our study: POSEMO and NEGEMO. These are weightings given to words in the analyzed text based on each word's positive or negative influence according to the LIWC dictionary. LIWC analysis of our corpus obtained the following results: it classified NEGATIVE entries 58.45% correctly, and POSITIVE entries, 88.10% correctly, for a total accuracy of 73.68%.

B. *Using LIWC's DEATH Category*

Human analysis of the incorrect LIWC's prediction for NEGATIVE entries noted that the system was particularly weak in classifying death-related events. These death-related events did not seem to raise the NEGEMO value. Fortunately, LIWC actually has a DEATH category that keeps track of words relevant to death-related events. A quick analysis revealed that if the DEATH value is greater than or equal to 1 as reported by LIWC, the entry is very likely to be NEGATIVE. With this addition, LIWC analysis classified NEGATIVE entries 83.80% correctly, and POSITIVE entries 87.41% correctly, for an overall classification accuracy of 85.61%. The results of LIWC's analysis are summarized in Figure 3.

Figure 3 The Percentage of Correct Classifications using LIWC POSEMO and NEGEMO Categories versus using POSEMO, NEGEMO, and DEATH categories



V. Combining the Statistical Learning Approach with the Hand-Crafted Database

To combine these two approaches, one could manually attempt to determine the relative weighting of each technique's scores. We chose to see if a data mining tool could effectively learn a decision tree based on the data. The See 5/C5.0 system, based originally on work by Ross Quinlan, was used to analyze the data and produce a decision tree [10, 11]. See5 builds decision trees from a set of training data using the concept of information entropy; the system divides data into smaller subsets by choosing a feature that maximizes information gain.
The inputs for the decision tree were as follows:

BayesianPositive: The raw score returned by the Bayesian analyzer indicating the probability the test entry is Positive.
BayesianNegative: The raw score returned by the Bayesian analyzer indicating the probability the test entry is Negative.
LIWCPositive: The POSEMO score returned by LIWC.
LIWCNegative: The NEGEMO score returned by LIWC.
LIWCDeath: The DEATH score returned by LIWC.
BayesianPosNegRatio: BayesianPositive/BayesianNegative.
LIWCPosNegRatio: POSEMO/ NEGEMO.

See5 was able to produce a decision tree that correctly classified 142 out of the 143 POSITIVE instances correctly (99.3%), and 139 out of the 142 NEGATIVE instances correctly (97.9%). The total accuracy is 98.6%. The decision tree produced by See5 for the entire database is given in Figure 4. The relative importance of each attribute can be seen in Table 3.

Figure 4 The Decision Tree Generated by See5 for the Classification of the 285 Texts

```
BayesianPosNegRatio <= 0.001466444:
:...LIWCPosNegRatio <= 1.997436: Negative (101)
:   LIWCPosNegRatio > 1.997436:
:   :...BayesianPositive <= 7.14e-025: Negative (28/1)
:       BayesianPositive > 7.14e-025: Positive (2)
BayesianPosNegRatio > 0.001466444:
:...LIWCNegative <= 2.5:
    :...LIWCPositive > 1.67: Positive (117)
    :   LIWCPositive <= 1.67:
    :   :...LIWCPosNegRatio > 14.12281: Positive (3)
    :       LIWCPosNegRatio <= 14.12281:
    :       :...LIWCDeath <= 0: Positive (13/3)
    :           LIWCDeath > 0: Negative (1)
    LIWCNegative > 2.5:
    :...BayesianPosNegRatio > 1087.901: Positive (6)
        BayesianPosNegRatio <= 1087.901:
        :...LIWCPosNegRatio > 1.996154: Positive (2)
            LIWCPosNegRatio <= 1.996154:
            :...BayesianPositive <= 1.97e-014: Negative (9)
                BayesianPositive > 1.97e-014:
                :...BayesianPositive <= 3.13e-007: Positive (2)
                    BayesianPositive > 3.13e-007: Negative (1)
```

Table 3 Percentage Usage of Each Attribute for Classification in the Decision Tree

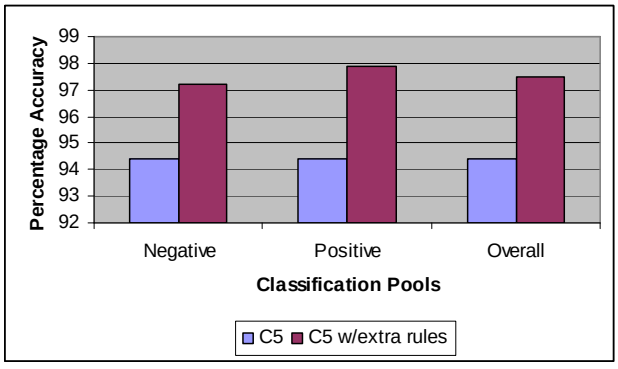
Attribute	% Usage
BayesianPosNegRatio	100
LIWCPosNegRatio	57
LIWCNegative	54
LIWCPositive	47
BayesianPositive	15
LIWCDeath	5

A. *Augmenting See5 with hand-built rules*

The tree presented in Figure 4 was generated by See5 using the entire database to create the decision tree. An additional experiment used leave-one-out training to train a decision tree and then predict the category for the "left out" data. Using this approach, the data was classified as follows: NEGATIVE texts were classified correctly 94.4%, and POSITIVE texts were classified correctly 94.4%, for an overall accuracy (perhaps obviously) of 94.4%.

Human analysis revealed that the See5 system tended to prune away some rules that would produce much better performance. For instance, the LIWC's DEATH category seems to play a more important role than is learned by the See5 system. If the DEATH value is greater than 1.3, the text is always NEGATIVE. Further if both LIWC and the Bayesian system indicate the passage is NEGATIVE, it is NEGATIVE. If the Bayesian system has a very high (> 1) POSITIVE ranking, it is POSITIVE (regardless of the LIWC POSEMO and NEGEMO scores). If the LIWC system has a score greater than zero for POSEMO and a score of 0 for NEGEMO and the Bayesian system is not below a small threshold, the entry should be POSITIVE. Using the decision tree produced by See5 augmented with the rules above, the integrated system produced the following results: For NEGATIVE entries, the accuracy was 97.2%; for POSITIVE entries, the accuracy was 97.9%; for an overall accuracy of 97.5%. A summary of these results is presented in Figure 5.

Figure 5 The Percentage of Correct Classifications using See5 versus See5 Augmented with Hand-Built Rules



VI. Future Directions

The main extension to this line of research will be to use more specific emotional categories. The standard categories specified by Ekman and used by the FACS coding system are fear, anger, sadness, happiness, disgust, and surprise [12]. In order to complete a similar analysis with these emotions, a new corpus of text will have to be created. Our ultimate goal is a system that synthesizes information from both spoken conversation and facial gesture. Thus, our corpus collection efforts will involve eliciting emotional spoken responses from videotaped subjects. Our likely mood induction method will be to show a highly evocative video and have the subjects describe their reaction to it. An interesting research question is what is the appropriate weighting to give to the text analyzer versus the facial gesture analyzer. Which is more important the words spoken or the expression on the speaker's face?

References

[1] Picard, R. W. *Affective Computing*, MIT Press, Cambridge, MA, 1997.

[2] Ratliff, M. and E. Patterson, Emotion Recognition Using Facial Expressions with Active Appearance Models, *Proceedings of the IASTED International Conference on Human-Computer Interface*, Innsbruck, March, 2008.

[3] Guinn, C. and R. Hubal. Extracting Emotional Information from the Text of Spoken Dialog, *9th International Conference on User Modeling*, Johnstown, PA. 2003.

[4] Wirth, R. J., Lecci, L., Denning, Crystal L., Little, Lindsay L., & Robertson, A. Examining the effects of priming on a mood induction procedure. *12th Annual Meeting of the American Psychological Society*, Miami, FL. 2000.

[5] Coan, J. and J. Allen, *Handbook of Emotion Elicitation and Assessment*, Oxford University Press, 2007.

[6] D. Jurafsky and J. Martin. *Speech and Language Processing*. Prentice-Hall, 2000.

[7] Guinn, C., Crist, D, and Werth, H., A Comparison of Hand-Crafted Semantic Grammars Versus Statistical Natural Language Parsing in Domain-Specific Voice Transcription, *Proceedings of Computational Intelligence*, Ed. B. Kovalerchuk, San Francisco, CA, pp. 490-495. 2006.

[8] Pennebaker, J., Francis, M., and R. Booth, http://www.liwc.net/. 2009.

[9] Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., & Booth, R. J. *The development and psychometric properties of LIWC2007*. Austin, TX, LIWC.Net. 2007.

[10] Quinlan, R. http://www.rulequest.com/. 2009.

[11] Quinlan, J. R. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, 1993.

[12] Ekman, P. and W. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*, Consulting Psychologists Press, Palo, Alto, 1978.