Predicting Success and Failure in Weight Loss Blogs through Natural Language Use

Cindy K. Chung[1] Clinton Jones[2] Alexander Liu[2] James W. Pennebaker[1]

The University of Texas at Austin
[1]Department of Psychology, 1 University Station Stop A8000, Austin, TX, 78712
[2]Department of Electrical and Computer Engineering, 1 University Station Stop C0803, Austin, TX, 78712
cindyk.chung@mail.utexas.edu, clint@idiotjones.com, ayliu@mail.utexas.edu, pennebaker@mail.utexas.edu

Abstract

We explore the emerging phenomenon of blogging about personal goals, and demonstrate how natural language processing tools can be used to uncover psychologically meaningful constructs in blogs. We describe features of a blog community (2638 blogs) devoted to weight loss. We compare several approaches to text analysis in predicting weight loss from natural language use in a subset of the blogs (258 users; over 13,000 entries). First, we use a bag of words approach to distinguish the degree to which individual words can predict success and failure. Next, we compare the results to a deductive word count and categorization tool, Linguistic Inquiry and Word Count. We discuss the theoretical significance of the words and word categories that distinguish between bloggers who succeed and those who fail in their weight loss attempts, along with the implications of automated text analysis in summarizing psychological features of blogs.

Blogging about Personal Goals

An emerging phenomenon has been the use of blogs for monitoring and documenting the process of self-change (e.g. Harmon, 2003). Thousands of bloggers are chronicling their daily successes and struggles with goals related to weight loss or debt. With each entry, information on weight lost or gained, or on money spent or saved is documented.

Examining language use in blogs presents an ecologically valid and unobtrusive way to bypass response rate and selection biases that result when online users are invited to participate in a research survey (see Lyons, Mehl, and Pennebaker 2006). Using text analytic tools, the words that people naturally use to write about the self-change process will be examined. The primary aims of this paper are 1) to find linguistic markers of self-regulatory success and failure in blogs devoted to weight loss, and 2) to compare two text analytic approaches in producing a good predictor model of weight loss success and failure in blogs.

Methods

Weight Loss Blogs Corpus

Publicly available blogs within a community of blogs devoted to weight loss (www.dietdiaries.com) were downloaded, resulting in 2638 blogs. Most bloggers updated their weight with each entry. The community had begun in 1999; the corpus was harvested in November 2007.

Text Samples

In order to sample from bloggers who had been committed to blogging about weight loss beyond a couple posts, and to sample many words per blog, a subset of blogs was selected for analysis. Blogs had to a) have posts over at least 4 months, b) have at least 25 posts, c) have updated weight information for the majority of posts, and d) the sex and goal weight of the blogger had to be apparent within the blog for manual annotations. In addition, obvious 2nd attempts by the same user were excluded. Due to the small number of male bloggers (only 10), blogs by males were excluded. These selection criteria resulted in a total of 258 blogs, for which the first 4 months of posts were downloaded, representing over 13,000 entries, and 3,243,695 words. The average word count of a blog in the subset was 12,572.5 words (*SD* = 10,472.3). Age, sex, and goal weight information were manually annotated.

On average, bloggers were 31.6 years old (*SD* = 12.0; exact age information available for 189 of the bloggers in the subset). The average starting weight of the bloggers in the subset was 208.5 lbs (*SD* = 53.6), and the average goal weight was 142.6 lbs (*SD* = 23.4), representing an average desired decrease of 65.9 lbs (*SD* = 45.9).

Percent of body weight loss over the first four months of blogging was computed by the following: (weight at 4 months – starting weight)/(starting weight). Positive numbers represented weight gain (and, accordingly, weight loss failure); negative numbers represented weight loss success. At the end of the 4 month period, bloggers had

lost an average of 13.8 lbs (*SD* = 12.9), representing an average body weight change of -6.53% (*SD* = 6.19%).

Analyses

Bag of Words Approach. In this approach, we treated the problem as a standard text classification task in machine learning. Classes were created by grouping users into groups based on amount of weight loss; we tried several different weight thresholds, resulting in a range of classes (2 to 5) with varying levels of balance between the class priors (ranging from very balanced to highly imbalanced). A bag-of-words model was then created after applying stop word removal, stemming, and TF-IDF weighting. A multinomial naïve Bayes classifier was then trained using ten random, stratified splits of the data set into training and test sets.

Linguistic Inquiry and Word Count (LIWC). LIWC (Pennebaker, Booth, and Francis 2007) is a computerized tool that counts the frequency of words and word stems in standard linguistic categories (e.g. articles, pronouns), psychological categories (e.g. emotion, and biological words), and various content categories (e.g. home, religion). Results are reported as a percentage of words in a given text file. For this study, LIWC category percentages indicated how much a given category was mentioned in the first 4 months of blogging. Correlations between LIWC categories and weight change were computed, and the significant correlations were entered into a regression equation. A prediction model and the variance accounted for in weight change were considered.

Results

Bag of Words Approach

Interestingly, the multinomial naïve Bayes classifier trained on the data was unable to discriminate between bloggers in different classes. In particular, the naïve Bayes classifier typically "learned" to almost always predict the largest class. In comparison, results using LIWC (which will be discussed below) had much more success in predicting weight loss.

LIWC

Percent of body weight change was correlated with sadness words (e.g. cry, sad), r_{258} = -.13, p = .03, and with ingestion words (e.g. carbs, eat), r_{258} = .17, p = .007. Percent of body weight change was unrelated to positive emotion words (e.g. awesome, happy), health words (e.g. nausea, sick), or social words (e.g. friend, hug).

Use of sadness and ingestion words were significant predictors of percentage of body weight change in a regression model, $F_{(2, 255)} = 5.13$, $p = .007$, $R^2 = .04$:

Change in body weight = -.06 - .04 (sadness) + .01 (ingestion)

Discussion

Weight loss success was significantly predicted by a greater use of sadness words, and fewer ingestion words. These results seem to suggest that sharing negative emotions is a more successful strategy in blogging about weight loss than simply keeping a food intake diary.

The results of our two text analytic approaches seem to be a case similar to tasks such as automatic author or gender identification (de vel et al. 2002) where features more suited to the task at hand are able to outperform a standard bag-of-words approach. In this case, it is better to look at classes of words (i.e., emotion words) rather than specific words like "muffin" which are used by all classes. The convenience of the LIWC software is that it provides a set of features consisting of previously validated word categories that have been found to be reliably associated with psychological constructs (Pennebaker, Mehl, and Niederhoffer 2003).

Using text analytic approaches, we were able to assess the linguistic correlates of successful weight loss, and to present a way for psychologists, market researchers, and users to understand the self-change process.

Acknowledgements: We would like to thank Joydeep Ghosh for his comments on the paper.

References

de Vel, O., Corney, M., Anderson, A., and Mohay, G. 2002. Language and gender author cohort analysis of e-mail for computer forensics. In *Proceedings Digital Forensics Research Workshop*, Syracuse, NY, USA.

Harmon, A. 2003. Finding comfort in strangers with an online diet journal. *The New York Times*. Retrieved October 7, 2007. http://query.nytimes.com/gst/fullpage.html?res=9A06E2DD1239F936A1575BC0A9659C8B63

Lyons, E. J., Mehl, M. R., and Pennebaker, J. W. 2006. Pro-anorexics and recovering anorexics differ in their linguistic Internet self-presentation. *Journal of Psychosomatic Research* 60: 253-256.

Pennebaker, J. W., Booth, R. J., and Francis, M. E. 2007. *Linguistic Inquiry and Word Count (LIWC) 2007: LIWC2007*. www.liwc.net; Austin, TX.

Pennebaker, J. W., Mehl, M. R., and Niederhoffer, K. 2003. Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology* 54: 547-577.