A Corpus-based Approach to Finding Happiness

Rada Mihalcea
Computer Science and Engineering
University of North Texas
rada@cs.unt.edu

Hugo Liu
Media Arts and Sciences
Massachusetts Institute of Technology
hugo@media.mit.edu

Abstract

What are the sources of happiness and sadness in everyday life? In this paper, we employ 'linguistic ethnography' to seek out where happiness lies in our everyday lives by considering a corpus of blogposts from the LiveJournal community annotated with happy and sad moods. By analyzing this corpus, we derive lists of happy and sad words and phrases annotated by their 'happiness factor.' Various semantic analyses performed with this wordlist reveal the happiness trajectory of a 24-day (3am and 9-10p are most happy), and a 7-day week (Wednesdays are saddest), and compare the socialness and human-centeredness of happy descriptions versus sad descriptions. We evaluate our corpus-based approach in a classification task and contrast our wordlist with emotionally-annotated wordlists produced by experimental focus groups. Having located happiness temporally and semantically within this corpus of everyday life, the paper concludes by offering a corpus-inspired livable recipe for happiness .

Finding Happiness

Whence happiness in everyday life? There are few questions that are as profound, as difficult to answer, or as timely and important for our stressful world. When we turn to great thinkers, we find disagreement—Aristotle equated happiness with virtue, Nietzsche found it in Greek tragedy and laughter, Thomas Jefferson saw it in human relationships. While the causes of happiness might never be fully universalized, each culture's *zeitgeist* should see trends. In this paper, we try to locate the sources of happiness for folks of the contemporary world by looking at culture's reflection in the blogposts of LiveJournal.com.

Albert Camus once threw caution to seekers of happiness—"You will never be happy if you continue to search for what happiness consists of"—but apparently he was not aware of 'corpus-based approaches,' which we suggest that if practiced judiciously, constitutes a computational tool for linguistic ethnography. Starting from a corpus of blogposts annotated as either *happy* or *sad*, we apply statistical language modeling techniques to ascertain the happiest and saddest words, phrases, and topics. We examine how everyday happiness ebbs and flows according to time-of-day and day-of-week, and how happy moments and sad moments are expressed through different modes of description.

Psychologists have collected something resembling happiness annotations for words within experimental settings, but we suggest through evaluation that the publicity of those settings leads to very different conclusions than those that we obtained by modeling blogs—arguably more private, honest, and polemic than opinions voiced in a psychological focus group. The paper concludes with an audacious attempt to articulate a corpus-inspired 'recipe for happiness.'

Corpus-based Evaluation of Happy/Sad Blogposts

Our study is based on a collection of blogposts from LiveJournal.com, annotated with *happy* and *sad* moods. Every blog community practices a different genre of writing—from our experience, LiveJournal.com blogs more closely recount the goings-on and happenstance of everyday life than any other blog community.

The indication of the mood is optional when posting on LiveJournal, therefore the mood-annotated posts we are using are likely to reflect the true mood of the blog authors, since they were explicitly specified without particular coercion from the interface. The corpus consists of 10000 blogposts, out of which half are labeled as *happy* and half are labeled as *sad*. This is a subset of the corpus used in the experiments reported in (Mishne 2005). In a pre-processing step, we removed all SGML tags and kept only the body of the blogposts, which was then passed through a tokenizer. We also filtered out all length outliers by removing all blogposts outside the 100-8000 characters range. The average length of the blogposts in the final corpus is 175 words / entry. Two sample entries are shown in Table 1.

[happy] Well kids, I had an awesome birthday thanks to you. =D Just wanted to so thank you for coming and thanks for the gifts and junk. =) I have many pictures and I will post them later. hearts
[sad] Home alone for too many hours, all week long ... screaming child, headache, tears that just won't let themselves loose.... and now I've lost my wedding band. I hate this.

Table 1. Sample blogposts in the *happy/sad* corpus.

A first aspect of interest was to evaluate how well the two moods can be separated by their linguistic content, which can be an indication of how reliable the mood annotations are. We trained a Naïve Bayes text classifier over unigram features and evaluated the classification accuracy in a five-fold cross-validation experiment. The accuracy was measured at 79.13%, which is a significant improvement with respect to the 50% accuracy associated with the naïve baseline of using one mood assignment by default[1]. As it turns out, the annotations in this data set are clearly separable, and therefore we can use this corpus to learn the characteristics of the *happy* and *sad* moods as indicated in the blog entries.

A second aspect of interest, which had even higher importance for our study, was the identification of a *happiness-factor* associated with the most salient words in the corpus. Starting with the features identified as important by the Naïve Bayes classifier (a threshold of 0.3 was used in the feature selection process), we selected all those features that had a total corpus frequency higher than 150, and consequently calculate the *happiness factor* of a word as the ratio between the number of occurrences in the *happy* blogposts and the total frequency in the corpus. This resulted in a list of 446 words. Table 2 shows the top most *happy* and top most *sad* words, together with their *happiness factor*.

[happy]		*[sad]*	
yay	86.67	goodbye	18.81
shopping	79.56	hurt	17.39
awesome	79.71	tears	14.35
birthday	78.37	cried	11.39
lovely	77.39	upset	11.12
concert	74.85	sad	11.11
cool	73.72	cry	10.56
cute	73.20	died	10.07
lunch	73.02	lonely	9.50
books	73.02	crying	5.50

Table 2. The happiness-factor of the most happy/sad words, given as percents.

While in this paper we pose ratings of words informally as happiness-factors, the psychology and artificial intelligence literatures demonstrate related efforts to acquire and compute the affect of text. Pennebaker *et al.*'s LIWC (2001) and Bradley and Lang's ANEW (1999) are two affectively annotated and normalized wordlists derived from experimental focus groups. Liu *et al.*'s Emotus Ponens textual affect sensor (2003) used word-level affect, in combination with large-scale world semantic knowledge, to assess the affective qualities of text, and also to model the attitudes and opinions of individuals (Liu & Maes 2004). Mishne (2005) experimented with affect classification of LiveJournal blogs, using cross-training and an ontology of over 100 moods.

We compared our corpus-derived wordlist annotated by happiness-factor against the ANEW corpus (1999) that was derived by ratings in experimental focus groups. Mapping our 446-word list into ANEW's 1000+ words, 23% (101 words) of our words were recalled in ANEW; we did not attempt to lemmatize our words to improve recall because morphology contains affective information in ANEW, e.g. there are different entries and values for "love" and "loved." Affective annotations in ANEW are given as three numbers (1-10) describing three dimensions of affect given by the widely used PAD model (Mehrabian 1995), that is (P)leasure-displeasure, (A)rousal-nonarousal, (D)ominance-submissiveness. The advantage of PAD is that it unifies emotional ontologies – for example, "fear" is low-pleasure, high-arousal, and low-dominance – and allows nuanceful discernments, such as between "joy" (high-arousal) and "contentment" (lower-arousal).

We compared our happiness-factor to ANEW's P-A-D values using Pearson's correlation and found that happiness weakly correlated to pleasure (r=0.54) and dominance (r=0.44), and did not correlate at all to arousal (r=-0.01). Much of the differing judgments can be attributed to the observation that while ANEW's experimental focus groups elicited a word's ideal emotions in the absence of real-world situations, LiveJournal's blogs reflect the reality of a word's emotions. For example, consider the idealistic versus realistic assessments of the words given in Table 3.

[WORD]	*[pleasure]*	*[happiness]*
love	87.2	48.7
god	81.5	35.0
father	70.8	27.9
baby	82.2	47.0

Table 3. Comparing different assessments of the emotional valence of some words by ANEW's Pleasure dimension and our blog-derived happiness factor (both normalized to the range (0%, 100%)

ANEW's *clinical optimism* is also reflected in the fact that for the 101 words compared, ANEW's pleasureableness had an average of 60.5%, while our happiness-factor had an average of 45.5%.

Happy Moments

People do not often talk about happiness, and even when they do, they usually repeat socially narrated attitudes about happiness. It is rare that people explicitly indicate what truly makes them happy, without being concerned of the opinion of others on the matter. Moreover, there seem to be stereotypes or default analogies about what constitutes a happy event – love, affection, fun, success –

[1] We also ran other classifiers, Support Vector Machines and Rocchio, but no significant differences were observed.

often times referring to "wanted", as opposed to current happiness. We refer to this as public happiness, which, while having its own merit, does not always coincide with the private feelings of happiness – the facts that trigger interior current feelings of happiness.

Since blogs are very close to diaries, they represent an ideal data set for deriving private expressions of happiness. While we admit that blogs do not have the same level of privacy as the old key-locked diaries, blog entries are nonetheless much closer to what people think, as compared to other forms of public expression.

Using the corpus of blog entries annotated with *happy* and *sad* feelings, we were able to derive the happiness load associated with the most salient concepts in the corpus, and sometimes the findings were surprising.

In particular, as mentioned before, the correlations between the happiness factor according to the blog data and the annotations in the ANEW corpus are relatively weak, and several of the words with high happiness load in blogposts were not even present in the anew corpus. This suggests that there is indeed a difference between private and public "displays" of happiness.

For instance, a topic that is often talked about in relation to happiness is *love*. In fact, looking up the ANEW list of words, *love* ranks the third out of 1034 words, with a high 87.2 *pleasure* factor. Instead, the word *love* in our list is neutral, with a happiness factor of only 48.7. Moreover, a morphological variation of the same word, *loved*, is ranked similarly high in the ANEW list (86.4), but it is marked as a sad word in our list (28.1). The only happy variation of this word is *lovely*, which has a happiness factor of 77.3 (this word does not appear in the ANEW list). What these rankings suggest is that the stereotypical view of *love* (the *public* view) is that of a happy word (the ANEW ranking), but the more *private* view of this word is neutral, due to a balanced distribution of this concept in both happy and sad moments.

There are also several examples of words in our list that have a high happiness factor, and yet they are not even

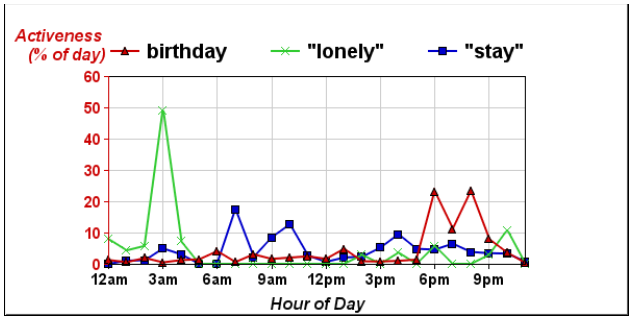

Fig. 1 Activeness-versus-time distributions derived for three words, using search engine statistics.

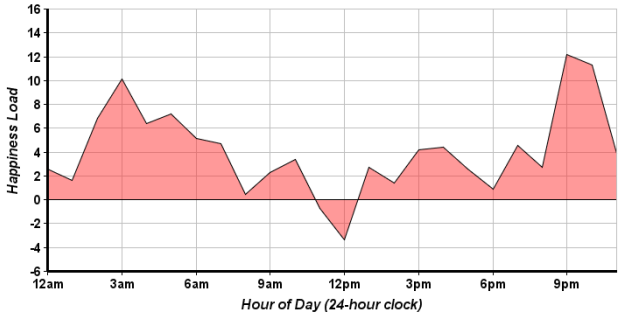

Fig. 2 The happiness load for a 24-hour day

listed in the ANEW list of words. For instance, *shopping* (79.5) and *bought* (69.9), are two of the highly ranked words in our list. Similarly, food related events (*lunch* (73.1), *dinner* (72.1) *food* (65.0), *drink* (64.4), etc.) have a high happiness factor, with only two of these concepts (*dinner* and *food*) being mentioned in the ANEW list. Another interesting example that has a positive happiness load in our list is *drunk* (59.1), which is most likely not one of the most widely publicly admitted states of happiness. Several other interesting cases observed in our corpus-derived collection of *private* happy moments are mentioned in the following sections.

The Time for Happiness

Happiness is certainly not a constant in our everyday life. There are times of the day when the happiness load is higher (e.g. *going out*), as compared to times with a lower happiness factor (e.g. *waking up*). An aspect of interest would therefore be the evaluation of the happiness load with respect to time (e.g. time of day, day of week), as determined by the activities that are typically performed at a certain point in time.

Time of Day

First, we determined the time distributions associated with the concepts in our list of salient words. For each word in the list, a set of 24 queries was run against the Web using a search engine, including each of the 24 hours in a day. For instance, *breakfast 9am* and *breakfast 1pm* are two of the queries ran for the concept *breakfast*. The counts collected in this way reflect the frequency of activities involving a certain concept, with respect to time. As an example, Figure 1 shows the time distributions of three words – *birthday*, *lonely*, and *stay*: *loneliness* peaks at 3am, while *birthday* is something that typically happens in the evening (6-9pm); the word *stay* instead is uniformly distributed over the entire 24-hour range, as there is no hour that is typical for activities associated with this word.

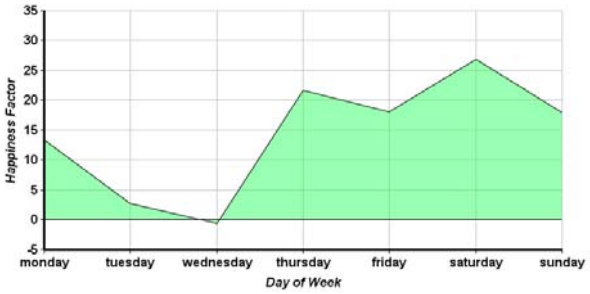

Fig. 3 The happiness factors for each day of the week.

24-hour time distributions were thusly obtained for each word on the 446-word list. Next, each distribution was weighted by its word's proportion of the total happiness pie (each factor was mapped to the range (-1,1) and divided through by the sum of the factors for all words) and all the weighted distributions were summed together to derive the happiness load of each of the 24 hours, as shown in Figure 2.

Waking up (7-9am, with a drop at 8am) appears to be an unhappy moment. Happiness increases as we advance through the day, with another drop at lunch time (food coma? sick of work already?). Things are again improving toward 4-5pm (time to go home), with a drop around 6pm (late day at work? traffic hours?), and then continuously increasing up until 10pm (night on the town). There is another peak of happiness around 2-4am, most likely due to the "night people" enjoying their night time, and perhaps betraying the biases of LiveJournal blogger lifestyles.

Day of Week

Another time scale that could potentially have an influence on the happy/sad feelings are the days of a week. Perhaps not surprising, all week days were part of our list of salient words as derived from the happy/sad corpus, and therefore we were able to use the happiness factor associated with each of the seven week days to plot the feelings of happiness across days of the week. Figure 3 shows the happiness factor for each of the seven days, remapped to a (-100%, 100%) scale for sake of consistency with Figure 2.

Thank God It's Friday (TGIF)? The results show that Saturday is actually the happiest day. According with intuition, the mid-week blues are sung on Wednesday, with an astronomical rise in happiness on Thursday, perhaps in anticipation of the weekend. Anticipation seems factored into each day's happiness, as Sunday is depressed by virtue of heralding in another long week, while Monday benefits from the happiness inertia of the recent weekend.

Happiness by Semantic Classes

As with temporality, we were also curious how happiness varied along various other semantic dimensions. In this section, we examine how three aspects of linguistic description (actions, things, attributes) and two content/topic dimensions (socialness, human-centeredness) predict happiness (see Figure 4).

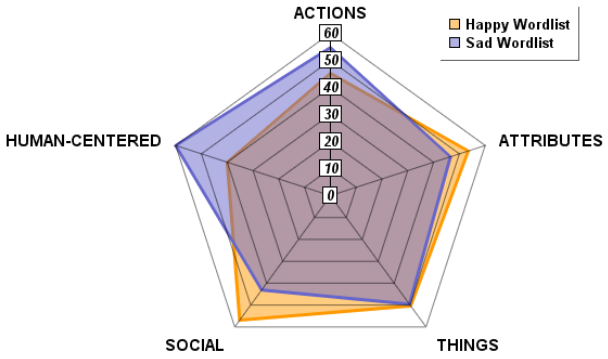

Fig. 4 Comparison of happy and sad blogposts along five semantic dimensions.

Modes of Description

By comparing the part-of-speech distributions of the happy-leaning and sad-leaning halves of the 446-word list, we found that happiness and sadness were conveyed using quite different modes of description.

For each word, all of its potential parts-of-speech are exposed. The "Actions" axis of Figure 4 is the percentage of words in each of the happy-leaning and sad-leaning lists that had "verb" as a possible part-of-speech. Likewise, "Things" corresponds to "nouns" and "Attributes" corresponds to "adjectives."

The implication of the results shown in Figure 4 is that sad descriptions rely more heavily on actions (e.g. "he lied," "I miss," etc.), while happy descriptions rely on more attributive and flowery descriptions (e.g. "new boyfriend," "great night"). To speculate, perhaps actions sadden folks and saddened folks take more actions. Perhaps happy folks do not act and remain contented, instead, they practice gushing flowery description on their current state.

Socialness

To assess the socialness of happiness and sadness, we fed the happy wordlist and sad wordlist into a major search engine. The number of results, taken as a measure of plausibility, was recorded for each word alone, and then for each word taken in the context of the socializing phrases "with friends," "with family," and "together." An example query looks like this – "*dinner with friends" OR "dinner with family" OR "dinner together."* Normalizing the number of social hits for a word by the total hits for the word alone – "*dinner*," we obtain roughly a metric of the word's socialness. Does the concept represented by the word occur in a social context, or rather is the concept reserved for the private self?

Comparing the average socialness of words in the happy and sad wordlists, we found that happy words were 32% more social than sad words. This finding accords with Thomas Jefferson's intuition that happiness is in the company of friends and family, and that being away from social company brings misery.

Human-centeredness

While sifting through the wordlist, we found that many of the happiest words seemed to concern objects (e.g. *books, pictures, food*) while the saddest words seemed invariably to concern much more existential matters such as close people and basic human activities (e.g. *father, mother, crying*). To test this hypothesis, we employed the categorization scheme of the WordNet (Fellbaum, 1998) lexical database to unify the words of our wordlist along the semantic dimension of human-centeredness. The happy and sad wordlists were lemmatized and mapped into WordNet with 72.8% and 81.6% recalls, respectively. The first and most frequent word sense is taken, and closure is performed over hypernym relations. If the word is subsumed by either the "act, human action, human activity," or the "psychological feature" top-level categories, then that word is judged as human-centered.

Employing this method, we found that sad words were 50% more human-centered than happy words. This result can be interpreted in two ways—human-centered drama (e.g. *father, mother*) causes much grief, or, sad descriptions often recount human acts of suffering (e.g. *cried, pain, loved*).

Note that the human-centeredness aspect refers to the concepts themselves which are classified as human-related or not. Human-centeredness can happen alone, while socialness implies the presence of other people. The fact that sadness is less social but more human-centered suggests that sadness is actually centered around one person—the author herself.

The Happiness of Facts

The happiness factor associated with single concepts can often be an indication of the happiness associated with the facts surrounding the given concept. For instance, a high happiness load associated with the word *lunch* is an indication that facts involving this concept would most likely have a high happiness load themselves. However, it is not always clear what these facts are, *have lunch* or *make lunch* or *after lunch*, and so on.

Although not facts *per se*, higher order n-grams can help us obtain a rough first approximation of the happiness factor associated with facts. Starting with the same corpus of blogposts as used in the unigram-based classification described before, we run a Naïve Bayes classifier on bigram and three-gram representations of the text. Not surprisingly, the classification accuracies obtained in five-fold cross-validations are similar to the unigram representation, i.e. 77.24% for the bigram representation, and 76.50% for three-grams. More interesting however is the ranking over the happiness factor associated with bigrams and three-grams, derived in the same way as for the unigrams (ratio between occurrences in the happy blogposts and the entire corpus, calculated for the features that are most salient according to the Naïve Bayes classifier). Tables 4 and 5 show the top most happy/sad bigrams and three-grams.

[happy]		*[sad]*	
happy happy	97.66	i wish	18.68
happy birthday	83.90	wish I	18.12
so happy	82.38	just want	16.84
my new	80.37	I hate	16.82
we went	76.98	miss you	14.79

Table 4. Most happy/sad bigrams

[happy]		*[sad]*	
happy happy happy	99.34	don't understand	10.44
all in all	88.67	I miss my	9.67
lot of fun	87.67	don't want to	8.33
happy birthday to	87.50	I'm sad	8.00
met up with	85.45	can't even	7.14

Table 5. Most happy/sad three-grams

An analysis of the happiness factor associated with bi-grams and three-grams reveals several interesting facts. For instance, *birthday* or phrases associated with this concept, such as *happy birthday*, is constantly ranked in the top of the list. Despite the typical use of this phrase as a wish to others, looking at the occurrences of the phrase in the corpus of blogposts shows that the vast majority refer to the birthday of the blog author, blogging about birthday wishes she received, which justifies the (typically self-centric) feeling of happiness.

Interestingly, a large number of the top-ranked *happy* bigrams refer to something new: *my new* (80.37), *the new* (67.29), *my first* (64.44), *a new* (61.85), which suggests that newness is one of the important factors resulting in happiness. This is in agreement with findings in experimental psychology that happiness in today's society occurs most often around novelty (Wilson, Meyers & Gilbert, 2001). The same study finds however, that happiness follows a trajectory of diminishing returns after the novel event has passed. Tim Wilson writes, "We don't realize how quickly we will adapt to a pleasurable event and make it the backdrop of our lives. When any event occurs to us, we make it ordinary. And through becoming ordinary, we lose our pleasure."

A large fraction of the bigrams with low happiness factor are centered around humans, i.e. *miss you, I hate, I wish, for him, you were*, etc. An observation in the same vein can be drawn from the ranking of three-grams, where facts revolving around *love* are related to humans in the top sad three-grams, *I love him* (37.33), *I love you* (37.09), but refer to objects in the three-grams with high happiness factor: *I love it* (85.00), *I love the* (82.85), *I love that*

(68.62). This reinforces the findings from the previous section, which showed that expressions of sad feelings are more often human-centered than the happy feelings.

Final Remarks

While most of the previous evaluations of the happiness load of concepts were derived through explicit human annotations, the goal of this study was to find the happiness associated with words and facts as it results from the *natural unconstrained* expression of feelings found in diary-like blog entries. As pointed out earlier, this unconstrained and more private type of writing allows us to identify expressions of happiness that occur *naturally*, and which reflect *private* feelings of happiness. Private happiness may sometimes differ from the kind of *public* happiness typically expressed in social circles—the only kind capturable within experimental focus group settings, such as those used to produce the ANEW wordlist.

We determined the happiness-load of words as obtained from blogposts labeled by their authors as *happy* or *sad*. These corpus-derived happiness-factors were then used as a basis for studies centered around the topic of happiness, which sometimes led to findings that contradict popular beliefs about the nature of happiness and sadness. We looked at the distribution of happiness across time (time of day and day of week), which revealed that certain hours and weekdays have higher happiness content than others. We also evaluated how five different semantic classes—actions, things, attributes, socialness, and human-centeredness—correlated with happy descriptions versus sad descriptions. Finally we looked at features derived from a classification task using higher order n-grams, and used the happiness-factor associated with bigrams and three-grams as an approximation for the happiness load of more complex facts (as compared to individual concepts). In addition to findings that enabled us to pinpoint the concepts and facts that bring happiness in our everyday life, the results of this study also allow us to suggest the following corpus-inspired recipe for happiness.

Recipe for Happiness[2]

Ingredients

- Something *new*
- *Lots* of *food* that you enjoy
- Your *favorite drink*
- An *interesting social* place

Directions

Go *shop* for something *new* – something *cool*, make sure that you *love it*. Then have *lots* of *food*, for *dinner* preferably, as the times of breakfast and lunch are to be avoided. Consider also including a *new*, *hot* taste, and one of your *favorite drinks*. Then go to an *interesting* place, it could be a *movie*, a *concert*, a *party*, or any other *social* place. Having *fun*, and optionally getting *drunk*, is also part of the recipe. Note that you should avoid any unnecessary actions, as they can occasionally trigger feelings of unhappiness. Ideally the recipe should be served on a *Saturday*, for maximum happiness effect. If all this happens on your *birthday*, even better. Bon appétit!

And so concludes this paper. *Yay*!

[2] Ingredients with high happiness load are italicized.

References

Bradley, M.M., & Lang, P.J. (1999). *Affective norms for English words (ANEW): Instruction manual and affective ratings.* Technical Report C-1, The Center for Research in Psychophysiology, University of Florida.

Fellbaum, C. WordNet: An electronic lexical database. MIT Press, Cambridge, MA, USA, (1998).

Hugo Liu, Henry Lieberman, Ted Selker. (2003). A Model of Textual Affect Sensing using Real-World Knowledge. Proceedings of the 2003 International Conference on Intelligent User Interfaces, IUI 2003, 125-132. ACM Press.

Hugo Liu and Pattie Maes. (2004). What Would They Think? A Computational Model of Attitudes. Proceedings of the ACM International Conference on Intelligent User Interfaces, IUI 2004, 38-45, ACM Press.

Mehrabian, A. (1995). Framework for a comprehensive description and measurement of emotional states. Genetic, Social, and General Psychology Monographs, 121, 339--361.

.

Mishne, G. 2005, Experiments with Mood Classification in Blog Posts, in Proceedings of the 1st Workshop on Stylistic Analysis Of Text For Information Access (Style 2005), Brazile, 2005.

Pennebaker, J.W., Francis, M.E., & Booth, R.J. (2001). Linguistic Inquiry and Word Count (LIWC): LIWC2001 (this includes the manual only). Mahwah, NJ: Erlbaum Publishers

Wilson, T. D., Meyers, J., & Gilbert, D. T. (2001). Lessons from the past: Do people learn from experience that emotional reactions are short lived? *Personality and Social Psychology Bulletin*, 27, 1648-1661.