Praiseworthy Act Recognition
Using Web-based Knowledge and Semantic Categories

Rafal Rzepka, Kohei Matsumoto and Kenji Araki
Graduate School of Information Science and Technology
Hokkaido University, Japan
{rzepka,matsumoto,araki}@ist.hokudai.ac.jp

Abstract

In this paper we[1] introduce our novel method for utilizing web mining and semantic categories for determining automatically if a given act is worth praising or not. We report how existing lexicons used in affective analysis and ethical judgement can be combined for generating useful queries for knowledge retrieval from a 5.5 billion word blog corpus. We also present how semantic categorization helped the proposed method to finally achieve 94% of agreement with human subjects who decided which act, behavior or state should be praised. We also discuss how our preliminary findings might lead to developing an important social skill of a robotic companion or an automatic therapist during their daily interaction with children, elderly or depressed users.

1 Introduction

Predictions from world demographic trends show that the current ratio of people aged sixty or more (12.6%) will nearly double in 2050 (almost 22%)[2]. Younger generations would need to work more and worry more, not only about their aged parents but also about their children to whom they would dedicate less time. Stress among working age group could be caused not only by work itself but also by the awareness of children and parents often left to their own devices. Data gathered by American Depression and Bipolar Support Alliance[3] indicates that depression most often strikes at age 32 in the United States, but poses also an obvious problem among different age groups. One child in 33 children and one in eight adolescents have clinical depression and even if as many as six million elderly people are affected by mood disorders, but only 10% ever receive treatment. Precise numbers are often difficult to obtain as many subjects do not want to participate in studies, do not respond to surveys, do not answer the door or have insufficient language abilities[4]. Problems related to psychological disorders could be alleviated by technological advancements, including progress in Artificial Intelligence, especially in cases of social withdrawal in which depressed adolescents prefer to deal with computers than with people. As psychology studies show [Hofmann *et al.*, 2012], the depression can be treated by cognitive behavioral therapies (CBT) as efficiently as medicaments and such treatment is based on conversation. Although computers are already used as supportive tools in CBT [Wright *et al.*, 2005], we are far away from entrusting patients to autonomous therapists. However, we believe that various conversational rules utilized in dialog-based therapies and other positive aspects [Burnard, 2003; Zimmerman *et al.*, 2009] of a conversation itself can be implemented in artificial agents like companion robots [Sarma *et al.*, 2014]. In this paper we introduce our idea how to utilize Natural Language Processing techniques, a set of lexicons and semantic categories to web mine knowledge necessary for recognizing if an action being a dialog topic should be e.g. complimented by an agent.

1.1 Importance of Praising

We chose the act of praising to be implemented in our artificial agent for a variety of reasons. First of all it is an evaluation task which positively influences a praised person [Kanouse *et al.*, 1981] and motivates, especially children [Henderlong and Lepper, 2002]. Often seen in interpersonal interaction, praising is used to encourage others, to socialize, to integrate groups, and to influence people [Lipnevich and Smith, 2008]. It is believed to have beneficial effects on self-esteem, motivation and performance [Weiner *et al.*, 1972; Bandura, 1977; Koestner *et al.*, 1987]. It is widely acknowledged that to praise oneself could substantially help dealing with depression [Swann *et al.*, 1992] and praising improves behavior [Garland *et al.*, 2008], academic performance [Strain *et al.*, 1983] and work performance [Crowell *et al.*, 1988]. But there is some other interesting and difficult aspect of praising – the praiser has to be competent and share some relationships with the praised person [Carton, 1996]. Also, from the Artificial Intelligence point of view, the automatic distinction between praiseworthy and not praiseworthy

[1] Second author is currently with Panasonic Co.
[2] www.unfpa.org/ageing
[3] www.dbsalliance.org
[4] www.nimh.nih.gov/health/statistics/prevalence/major-depression-among-adults.shtml

Proceedings of the 4th Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2016), IJCAI 2016, pages 41-47, New York City, USA, July 10, 2016.

acts is an interesting long-term challenge to create a righteous and trustful machine and, in this particular case, to investigate if the Web resources could become a sufficient knowledge base for such tasks. Our hypothesis is that knowing the polarity of consequences of human acts might be the key to an automatic evaluation of these acts.

1.2 State of the Art

The authors have found only one research proposal dedicated specifically to automating praising. In 1998 [Tejima *et al.*, 1998] have published a two page paper in which they describe their observations from physiotherapists' sessions with elderly. The researchers proposed a simple verbal encouragement algorithm for walking training and implemented it later [Tejima and Bunki, 2001], however the effectiveness could not be confirmed due to the insufficient number of experimental subjects. Causing positive moods in interlocutors can be found as a sub-task in Human-Computer Interaction (HCI) field, especially in learning-oriented agents [Fogg and Nass, 1997; Kaptein *et al.*, 2010] but the studies utilize scenarios and manually created rules when to praise. Systems that accept, in theory, any sentence as an input and recognize polarity or emotive categories were proposed in the fields of sentiment analysis and affect recognition [Wilson *et al.*, 2005; Strapparava and Mihalcea, 2008] and the basic idea for our system is borrowed from their approaches. However these methods cannot be utilized straightforwardly because *being positive* does not have to mean an act is *worth praising* ("I saw a movie" is labelled positive by these methods but it usually does not mean we need to react with a compliment to such a statement). For English language there are promising methods for retrieving *goodFor* and *badFor* events [Deng and Wiebe, 2014] and for acquiring knowledge of stereotypically positive and negative events from personal blogs [Ding and Riloff, 2016]. Basically any new trend in the field [Cambria *et al.*, 2013; Socher *et al.*, 2013] should eventually help improve our results as soon as they are implemented for Japanese language, which often has much less resources to keep up with the latest methods. For Japanese [Rzepka and Araki, 2015] have proposed a system that evaluates textual inputs from a moral perspective. Similarly to our approach it uses lexicons and one of them, based on Kohlberg's theory of moral stages development [Kohlberg, 1981], includes praise-punishment polarized pairs. However, the lexicon contains only 14 praise related words limited to synonyms of the verb "praise" which, as shown later in the comparison experiment, are insufficient for our purposes.

2 System Overview

The algorithm of our system is presented in Figure 1. In the first step an input act (noun - verb pair we treat as the minimal semantic unit describing any act) in Japanese language is morphologically analyzed by MeCab[5] to determine a noun, a verb and the joining particle representing grammatical case (e.g. *aisatsu-o suru* "to greet someone" or *yakusoku-o mamoranai* "not keeping promises", where particle "o" indicates an object of given verb). Then the system adds to

[5] taku910.github.io/mecab/

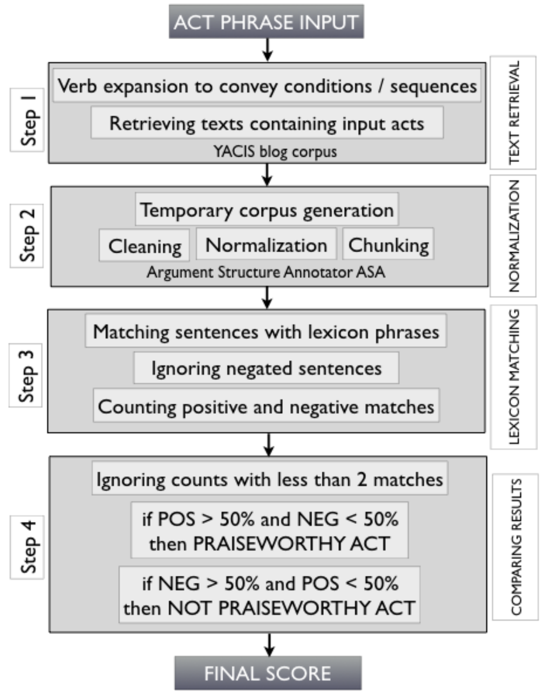

Figure 1: Algorithm for retrieving and analyzing consequences of acts in order to determine if they should be praised.

the verb 15 suffixes representing conditions and temporal sequences to retrieve more adequate sentences (*waruguchi itta ato* "after calling names", *waruguchi iu toki* "when called names", *waruguchi itte* "called names and then", etc.). Because particles are often omitted in colloquial Japanese, another set of 15 phrases without particles is created and the final 30 phrases together with phrases with verbs in their basic (dictionary) form become queries for 5.5 billion word YACIS corpus of Japanese raw blogs [Ptaszynski *et al.*, 2012]. Text retrieved from the corpus is then cleaned – emoticons usually used as sentence boundaries are converted to fullstops and too long and too short sentence candidates are deleted. In the next step, the generated temporary corpus of sentences containing input acts is normalized to verb dictionary forms and divided into meaningful chunks by Argument Structure Annotator ASA [Takeuchi *et al.*, 2010] to avoid granular division of morphological analyzer. For instance "was | beat | ing | brother" becomes "beat brother" and such transitions are made to increase the coverage of matching chunks with phrases from lexicons in the next step. Every match is scored 1 and the totals are compared. If there are more than 50% of positive or negative counts, the act is estimated as praiseworthy or not praiseworthy accordingly. Although in morality estimation task 60/40 ratio scored highest [Rzepka and Araki, 2015], in our task the 50/50 ratio achieved better results.

2.1 Lexicons

As mentioned in the introduction, we hypothesized that measuring the polarity of act consequences might be the key for recognizing praiseworthy acts. Although aware of possible problems mentioned in the Introduction, we decided to investigate how efficient the existing emotional recognition methods could deal with our task. Therefore firstly we chose two different freely available lexicons used for lexicon based polarity recognition in Japanese language. The larger one was statistically generated from manually annotated sentences in the study of [Takamura *et al.*, 2005]. It contains 55,102 words divided into positive (5,121 words) and negative (49,981 words) ones. Every word was automatically scored on the scale from minimal -1 to maximal 1 and the words closer to 0 tend to be inaccurately labeled (e.g. *okaasan* “mom” or *narubeku* “as possible”, are marked as negative words), therefore using the whole (significantly unbalanced) set would cause drops in accuracy. In order to minimize this problem and to make the lexicon more balanced, after analyzing the entries we used most positive 3,000 and most negative 3,000 words (closest to 1 and -1 from each side) and called it “Statistical Lexicon”.

Another lexicon used in polarity detection in Japanese texts is created manually by [Nakamura, 1993] from emotive sentences retrieved from Japanese literature. The words are separated into ten categories (Like, Joy, Relief, Dislike, Anger, Fear, Shame, Sadness, Excitement, Surprise) and because Excitement and Surprise have no distinct valence, these two categories were excluded. The combined words from Like, Joy and Relief form a positive subset and Dislike, Anger, Fear, Shame and Sadness form a negative one. Resulting lexicon of 526 positive and 756 negative words (1,282 in total) we call here “Literature Lexicon” to make it more comprehensible while presenting comparison between lexicons.

As mentioned before, a positive act does not necessarily imply being praiseworthy, therefore we decided also to test a lexicon used for ethical judgement by [Rzepka and Araki, 2015]. This relatively small set, containing 65 positive and 69 negative words (134 in total), was created by applying phrases related to the five stages of moral development proposed by [Kohlberg, 1981]: obedience / punishment, self-interest, social norms, authority / social-order, and social contract. For example in the obedience / punishment subset there are words like “punished”, “awarded”, “punishment”, “award” and authority / social order contains law-related words like “sentenced”, “legal” or “arrested”. To examine how emotional and social consequences work together, we created another lexicon, a combination of Kohlberg’s theory-based set with the Nakamura’s literature-based set. We named the former “Ethical Lexicon”, and the latter “Combined Lexicon”.

3 Experiments and Results

In this section we introduce experiments we conducted to investigate the effectiveness of our approach in the task of automatic praiseworthy act recognition.

3.1 Input Acts

Web resources used in the study give an opportunity to process any kind of act but this freedom causes difficulties with choosing a fair and balanced input. To deal with this problem we created two sets, one generated automatically and evaluated by subjects, and second one created by the same subjects specifically instructed to give examples of praiseworthy and not praiseworthy acts different from these which they labeled. By introducing these two types we tried to find a balance between “any input” (because the algorithm should recognize neutral acts) and more specific, manually crafted set of correct data.

Automatically Generated Set

For creating the first set we utilized 200 verbs from the Statistical Lexicon with the highest hit number in the blog corpus (100 from positive subset and 100 from negative subset) and paired them with nouns most frequently co-occurring within Japanese Frames dataset automatically generated from the biggest Japanese Web corpus [Kawahara and Kurohashi, 2006]. In order to limit the number of acts and to maintain sufficient coverage (to observe to what extent the automatically polarized words are efficient), we added two conditions. The noun object must be included in the Statistical Lexicon and the generated act must appear at least ten times in the blog corpus. Hence, if e.g. verb “keep” from the lexicon was co-occurring frequently with object noun “promise” and the phrase “to keep a promise” was found more than 10 times in the blog corpus, the phrase was treated as a common human act and became an input. With this method we generated 119 acts which were then evaluated by three judges (one female in her fifties, one male university student and one female secondary school pupil) by labeling the set as *praiseworthy*, *not praiseworthy* or *hard to tell*. The majority vote (three judges agreed or two agreed and the third answered “hard to tell”) resulted in 54 acts – 31 worth praising as *tomodachi-o iwau* (“to congratulate a friend”) or *chichi-o shitau* (“to admire one’s father”) and 23 not worth praising as *tanin-o nikumu* (“to hate somebody”) or *itami-o shiiru* (“to impose pain upon someone”). Two examples of acts on which agreement was not reached are *hiza-o kussuru* (“to bend one’s knees / to yield to someone”) and *yami-o kowagaru* (“to be afraid of darkness”). The labeled data became both the input and first correct data set and we named it “Automatically Generated Set”.

Manually Created Set

Because the automatically retrieved input set was biased toward Statistical Lexicon we asked the same group of three people to think of acts worth praising and not worth praising. The created set (from now on called “Manually Created Set”) contained 64 acts – 32 of praiseworthy ones as *shiken-ni goukaku suru* (“passing an exam”) or *tetsudai-o suru* (“helping someone”), and not worth praising as *yakusoku-o mamoranai* (“not to keep a promise”), *kenka-o suru* (“to quarrel / to have a fight”). Differently from the Automatically Generated Set, although the creators have seen examples of acts in the evaluation process, Manually Created Set was not restricted and in consequence included more diverse forms containing not only negations but also adverbs and passive /

Table 1: Results for Automatically Generated Set of input acts.

	Matched / All	Correct
Statistical Lexicon	54 / 54	83.3%
Literature Lexicon	42 / 54	66.7%
Ethical Lexicon	17 / 54	58.8%
Combined Lexicon	45 / 54	68.9%

Table 2: Results for the Manually Created Set of input acts.

	Matched / All	Correct
Statistical Lexicon	52 / 64	63.5%
Literature Lexicon	45 / 64	84.4%
Ethical Lexicon	39 / 64	84.6%
Combined Lexicon	44 / 64	90.9%

double verbs as in *jiko-chuushin-teki ni koudou-o suru* ("to act selfishly") and *iwareta koto-o yaranai* ("not to do what one was told").

3.2 Effectiveness Comparison between Lexicons

Having two sets of acts with their human evaluation prepared, we have performed a series of experiments to examine our system's accuracy when using above described lexicons in the task of recognizing praiseworthy acts.

Statistical Lexicon

Tested with acts from the Automatically Generated Set, the Statistical Lexicon achieved 83.3% of correct recognitions. To confirm our assumption that matching should be performed only on the right side of an act phrase because it is where consequences of the act are usually written (see Figure 2), we have also run additional tests and confirmed that analyzing left sides achieves significantly lower accuracy (66.7%). Matching within the whole sentence did not bring any improvement in results, besides it doubled searching time. Examples of correctly recognized acts are *shouri-o iwau* ("to celebrate victory") and *kenkou-o mamoru* ("to care about one's health"). On the other hand, *tsumi-o kuiru* ("to regret one's sins") or *shi-o kanashimu* ("to grieve one's death") were recognized incorrectly due to noisy polarity in the Statistical Lexicon.

When tested with Manually Created Set, the results of Statistical Lexicon dropped as expected. Left side matching brought only 53.7% correct recognitions while again the right side matching surpassed the left side achieving 63.5% and the whole sentences scored significantly lower (58.2%). All other comparison of results between left side, right side and whole sentences confirmed this trend, therefore, in order to avoid confusion, all remaining results we introduce, are from the matches performed on the right sides following input act phrases.

Literature Lexicon

The Literature Lexicon surpassed much larger Statistical Lexicon when Manually Created Set acts were input but was significantly less accurate with acts from Statistical Lexicon (see Table 1 and Table 2). The perfect recognition rate (54/54 matched) may suggests that if a new, less noisy method for the automatic estimation of word polarity is proposed and it covers all words in every possible input, the Statistical Lexicon would outperform the Literature Lexicon also when fed with acts from Manually Created Set. Nevertheless, it would be very costly and avoiding polarizing neutral words seems to be difficult, hence we believe that using manually crafted, small lexicons is currently more realistic approach for the automatic recognition (and annotation) of praiseworthy acts.

Ethical Lexicon

The smallest of all used lexicons, based on Kohlberg's theory and utilized in automatic ethical recognition task performed worst when the Automatically Generated Set of acts was input but outperformed both Statistical and Literature Lexicons when the Manually Created Set of acts was used.

Combined Lexicon

We managed to confirm that the combination of Ethical and Literature Lexicons performed better than separated ones when the Manually Created Set of acts was used. However, its accuracy was still lower than Statistical Lexicon matching sentences retrieved with the Automatically Created Set of acts.

3.3 Additional Experiments

As we aim at recognizing praiseworthy acts in everyday conversation, the correct recognition of more natural input acts is more important than the correct recognition of less natural input acts. To be sure if Statistical Lexicon could perform better with Manually Created Set we conducted a series of additional tests increasing the range of positive and negative words to see if heuristically chosen size of 3,000 was correct. We examined 10 sizes starting from 500 words size increasing it by 500 each time up to 5,000 and also tested the whole unbalanced list from -1 to 1. It appeared (See Figure 3) that accuracy grows till 1,500 words (increase from 72.9% to 80.8%) but when a larger sets are used, the results start to decrease and never exceed these of the Literature Lexicon (84.4%).

4 Adding Semantic Categories

After analyzing sentences which include praiseworthy act but were not counted due to insufficient number of words in lexicons we decided to examine if we could automatically add some valuable information to other words and see if the information influences the act of praising. We chose semantic categorization and used "Bunrui-Goi-hyo" (Word List by Semantic Principles) [NLRI, 1964] containing 32,600 semantically categorized words collected from 90 contemporary Japanese newspapers. For example the list groups words under categories as "Thoughts / Opinions / Doubts", "Helping / Rescuing" or "Profit / Loss". Our idea was to add simple weighs (count +1) to words that belong to categories which tend to be praiseworthy. In order to examine which categories reveal such tendencies we retrieved from the corpus all sentences containing acts labeled by human subjects

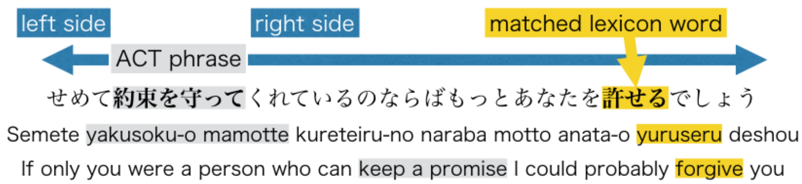

Figure 2: Example sentence from the corpus with input act and a matched Ethical Lexicon word on the right side.

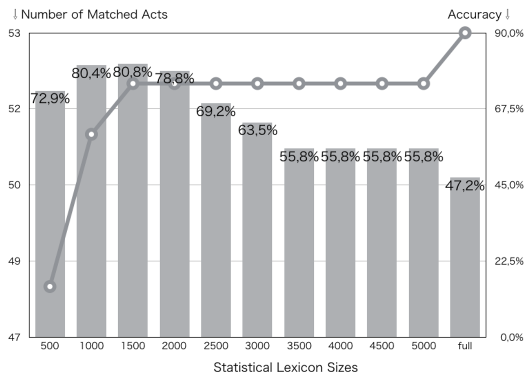

Figure 3: Results of additional experiments for investigating accuracy changes when using different sizes of the Statistical Lexicon.

as praiseworthy and not praiseworthy. Then a simple script counted how many other words in both datasets belong to which semantic category. For example if a blog sentence was "I lost the confidence in myself after he spoke ill about me", the script was adding negative points to categories as "Profit/Loss" (*lost*) or "Thoughts / Opinions / Doubts" (*confidence*). Because some categories contained thousands of words and other only a few, we decided to assign weights according to differences between frequencies. Examples of categories with distinctly different frequencies are shown in Table 3. Then, in order to ease unbalance between sizes of both categories, we experimented with combinations of weight sets and discovered that accuracy is highest for both praiseworthy and not praiseworthy acts when the former uses weights created from group b) and the latter uses c) (refer to Table 3).

4.1 Result Comparison

To see if semantic categorization is effective, we repeated all experiments scoring not only matched lexicon words but also other words that belong to specific categories (those with tendencies to be praiseworthy or not praiseworthy). Because among semantic categories supposedly specific to praiseworthy acts there were ones like Losing and Disappointment, we expected rather low accuracy, but quite surprisingly semantic weighting helped improving all previous results (see Table 4 and Table 5). Even when we excluded lexicon words count entirely, the semantic categories alone achieved slightly better precision than Ethical Lexicon when the Automatically Generated Set of acts was input. The highest precision when Manually Created Set was used increased the precision of Literature and Ethical Lexicons achieving 94%.

5 Conclusion, Future Work and Discussion

In this paper we introduced a simple matching algorithm allowing an agent to recognize human acts worth praising with maximal 94% agreement with human subjects by using lexicons (words sets) and Web resources (a blog corpus). The best results were achieved by Literature and Combined Lexicons with Semantic Categories support when manually created example acts were input. There is still plenty of room for improvement and we plan to increase the coverage of lexicons by matching synonyms, too. We also are experimenting with changing counting method according to adverbs preceding matching phrases ("a little bit sad" could be scored lower than e.g. "so freaking sad"). As the act of praising is very subjective and depends on many factors, we are planning to perform wide, possibly intercultural, surveys. We would like to conclude with underlining a wider importance of the ability to automatically recognize praiseworthy acts by a machine. Recent worries about Artificial Intelligence taking control over their users could be, at least in our opinion, eased by positive examples. Companion robots, while helping at home and e.g. running memory-quizes for users with Alzheimer disease, need to be trusted and gaining the trust will be difficult without sharing similar values. Our common recognition and evaluation of a fellow human's behavior can be measured with shallow sentiment analysis techniques on vast textual data which express our experiences and feelings. The proposed method demonstrates that the noisy Web resources like blogs, when processed carefully, can become one way to equip artificial agents with a human-like capacity of telling right from wrong without leaning to any specific philosophy or religion. We believe that a trustworthy machine should rather operate on estimating overall positive and negative consequences than on methods based on explicit rules decided by one or only few programmers. The proposed system can easily "explain" its decisions by giving examples of retrieved experiences or by presenting a voting ratio, while most of machine learning based methods are "black boxes" and may lead to trust issues. Having said so, we believe that our method could help to automatically annotate data, which is crucial for machine learning.

Table 3: Examples of frequency differences of semantic categories specific to praiseworthy and not praiseworthy acts

Difference	**Praiseworthy acts**
a) More than 4 times:	Helping / Rescuing, Giving / Receiving, Profit / Loss, Winning / Losing, School / Military, Lending / Borrowing, Physiology, Marking / Signing, etc.
b) More than 3 times:	Talents, Planning, Specialist jobs, Associations / Groups, Events / Ceremonies, etc.
c) More than 2 times:	Economy / Income / Expenditure, Formation, Meaning / Problem / Purpose, Desire / Expectance / Disappointment, etc.

Difference	**Not praiseworthy acts**
a) More than 4 times:	Respecting / Thanking / Trusting, Creating / Writing, Old / New / Slow / Fast, Treatment, Graphs / Tables / Formulas, etc.
b) More than 3 times:	Acquisition, Eye / Mouth / Nose functions, Roads / Bridges, Land vehicles, Fear / Anger, etc.
c) More than 2 times:	Linguistic activities, Birds, Associations, Distress / Sorrow, Partners / Colleagues, etc.

Table 4: Effectiveness comparison of implementing semantic categories (Automatically Generated Set).

	Matched / All	Correct
Semantic Category (SC)	52 / 54	78.8%
Statistical Lexicon + SC	54 / 54	85.2%
Literature Lexicon + SC	54 / 54	81.5%
Ethical Lexicon + SC	52 / 54	76.9%
Combined Lexicon + SC	54 / 54	85.2%

Table 5: Effectiveness comparison of implementing semantic categories (Manually Created Set).

	Matched / All	Correct
Semantic Category (SC)	50 / 64	92.0%
Statistical Lexicon + SC	52 / 64	88.5%
Literature Lexicon + SC	50 / 64	94.0%
Ethical Lexicon + SC	50 / 64	90.0%
Combined Lexicon + SC	50 / 64	94.0%

References

[Bandura, 1977] Albert Bandura. Self-efficacy: toward a unifying theory of behavioral change. *Psychological review*, 84(2):191, 1977.

[Burnard, 2003] Philip Burnard. Ordinary chat and therapeutic conversation: phatic communication and mental health nursing. *Journal of Psychiatric and Mental Health Nursing*, 10(6):678–682, 2003.

[Cambria *et al.*, 2013] E. Cambria, B. Schuller, Yunqing Xia, and C. Havasi. New avenues in opinion mining and sentiment analysis. *Intelligent Systems, IEEE*, 28(2):15–21, March 2013.

[Carton, 1996] John S Carton. The differential effects of tangible rewards and praise on intrinsic motivation: A comparison of cognitive evaluation theory and operant theory. *The Behavior Analyst*, 19(2):237, 1996.

[Crowell *et al.*, 1988] Charles R Crowell, D Chris Anderson, Dawn M Abel, and Joseph P Sergio. Task clarification, performance feedback, and social praise: Procedures for improving the customer service of bank tellers. *Journal of Applied Behavior Analysis*, 21(1):65–71, 1988.

[Deng and Wiebe, 2014] Lingjia Deng and Janyce Wiebe. Sentiment propagation via implicature constraints. In *EACL*, pages 377–385, 2014.

[Ding and Riloff, 2016] Haibo Ding and Ellen Riloff. Acquiring knowledge of affective events from blogs using label propagation. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)*, 2016.

[Fogg and Nass, 1997] B.J. Fogg and C. Nass. Silicon sycophants: the effects of computers that flatter. *International Journal of Human-Computer Studies*, 46(5):551 – 561, 1997.

[Garland *et al.*, 2008] Ann F Garland, Kristin M Hawley, Lauren Brookman-Frazee, and Michael S Hurlburt. Identifying common elements of evidence-based psychosocial treatments for children's disruptive behavior problems. *Journal of the American Academy of Child & Adolescent Psychiatry*, 47(5):505–514, 2008.

[Henderlong and Lepper, 2002] Jennifer Henderlong and Mark R Lepper. The effects of praise on children's intrinsic motivation: a review and synthesis. *Psychological bulletin*, 128(5):774, 2002.

[Hofmann *et al.*, 2012] Stefan G Hofmann, Anu Asnaani, Imke JJ Vonk, Alice T Sawyer, and Angela Fang. The efficacy of cognitive behavioral therapy: a review of meta-analyses. *Cognitive therapy and research*, 36(5):427–440, 2012.

[Kanouse *et al.*, 1981] David E Kanouse, Peter Gumpert, and Donnah Canavan-Gumpert. The semantics of praise. *New directions in attribution research*, 3:97–115, 1981.

[Kaptein *et al.*, 2010] Maurits Kaptein, Panos Markopoulos, Boris Ruyter, and Emile Aarts. Two acts of social intelligence: the effects of mimicry and social praise on the evaluation of an artificial agent. *AI & SOCIETY*, 26(3):261–273, 2010.

[Kawahara and Kurohashi, 2006] Daisuke Kawahara and Sadao Kurohashi. A fully-lexicalized probabilistic model for japanese syntactic and case structure analysis. In *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, HLT-NAACL '06, pages 176–183, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

[Koestner *et al.*, 1987] Richard Koestner, Miron Zuckerman, and Julia Koestner. Praise, involvement, and intrinsic motivation. *Journal of personality and social psychology*, 53(2):383, 1987.

[Kohlberg, 1981] Lawrence Kohlberg. *The Philosophy of Moral Development*. Harper and Row, 1th edition, 1981.

[Lipnevich and Smith, 2008] Anastasiya A Lipnevich and Jeffrey K Smith. Response to assessment feedback: The effects of grades, praise, and source of information. *ETS Research Report Series*, 2008(1):i–57, 2008.

[Nakamura, 1993] Akira Nakamura. *Kanjo hyogen jiten [Dictionary of Emotive Expressions]*. Tokyodo Publishing, 1993.

[NLRI, 1964] National Language Research Institute NLRI. *Bunrui Goi Hyo (Word List by Semantic Principles, in Japanese)*. Shuei Shuppan, 1964.

[Ptaszynski *et al.*, 2012] Michal Ptaszynski, Pawel Dybala, Rafal Rzepka, Kenji Araki, and Yoshio Momouchi. Yacis: A five-billion-word corpus of japanese blogs fully annotated with syntactic and affective information. In *Proceedings of The AISB/IACAP World Congress*, pages 40–49, 2012.

[Rzepka and Araki, 2015] Rafal Rzepka and Kenji Araki. *Rethinking Machine Ethics in the Age of Ubiquitous Technology*, chapter Semantic Analysis of Bloggers Experiences as a Knowledge Source of Average Human Morality, pages 73–95. Hershey: IGI Global, 2015.

[Sarma *et al.*, 2014] Bandita Sarma, Amitava Das, and Rodney D Nielsen. A framework for health behavior change using companionable robots. *INLG 2014*, page 103, 2014.

[Socher *et al.*, 2013] Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, volume 1631, page 1642. Citeseer, 2013.

[Strain *et al.*, 1983] Phillip S Strain, Deborah L Lambert, Mary Margaret Kerr, Vaughan Stagg, and Donna A Lenkner. Naturalistic assessment of children's compliance to teachers'requests and consequences for compliance. *Journal of Applied Behavior Analysis*, 16(2):243–249, 1983.

[Strapparava and Mihalcea, 2008] Carlo Strapparava and Rada Mihalcea. Learning to identify emotions in text. In *Proceedings of the 2008 ACM symposium on Applied computing*, pages 1556–1560. ACM, 2008.

[Swann *et al.*, 1992] William B Swann, Richard M Wenzlaff, and Romin W Tafarodi. Depression and the search for negative evaluations: more evidence of the role of self-verification strivings. *Journal of Abnormal Psychology*, 1992.

[Takamura *et al.*, 2005] Hiroya Takamura, Takashi Inui, and Manabu Okumura. Extracting semantic orientations of words using spin model. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 133–140. Association for Computational Linguistics, 2005.

[Takeuchi *et al.*, 2010] Koichi Takeuchi, Suguru Tsuchiyama, Masato Moriya, and Yuuki Moriyasu. Construction of argument structure analyzer toward searching same situations and actions. Technical Report 390, IEICE technical report. Natural language understanding and models of communication, jan 2010.

[Tejima and Bunki, 2001] Noriyuki Tejima and Hitomi Bunki. Feasibility of measuring the volition level in elderly patients when using audio encouragement during gait training physical therapy. In *Engineering in Medicine and Biology Society, 2001. Proceedings of the 23rd Annual International Conference of the IEEE*, volume 2, pages 1393–1395. IEEE, 2001.

[Tejima *et al.*, 1998] Noriyuki Tejima, Yoko Takahashi, and Hitomi Bunki. Verbal-encouragement algorithm in gait training for the elderly. In *Engineering in Medicine and Biology Society, 1998. Proceedings of the 20th Annual International Conference of the IEEE*, volume 5, pages 2724–2725. IEEE, 1998.

[Weiner *et al.*, 1972] Bernard Weiner, Heinz Heckhausen, and Wulf-Uwe Meyer. Causal ascriptions and achievement behavior: a conceptual analysis of effort and reanalysis of locus of control. *Journal of personality and social psychology*, 21(2):239, 1972.

[Wilson *et al.*, 2005] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 347–354. Association for Computational Linguistics, 2005.

[Wright *et al.*, 2005] Jesse H. Wright, Andrew S. Wright, Anne Marie Albano, Monica R. Basco, L. Jane Goldsmith, Troy Raffield, and Michael W. Otto. Computer-assisted cognitive therapy for depression: Maintaining efficacy while reducing therapist time. *The American Journal of Psychiatry*, 162(6):1158–64, Jun 2005.

[Zimmerman *et al.*, 2009] Frederick J Zimmerman, Jill Gilkerson, Jeffrey A Richards, Dimitri A Christakis, Dongxin Xu, Sharmistha Gray, and Umit Yapanel. Teaching by listening: The importance of adult-child conversations to language development. *Pediatrics*, 124(1):342–349, 2009.