Opinion Marks: A Human-Based Computation Approach to Instill Structure into Unstructured Text on the Web

Bum Chul Kwon
Universität Konstanz
bumchul.kwon@uni-konstanz.de

Jaegul Choo
Korea University
jchoo@korea.ac.kr

Sung-Hee Kim
Purdue University
kim731@purdue.edu

Daniel Keim
Universität Konstanz
keim@uni-konstanz.de

Haesun Park
Georgia Institute of Technology
hpark@cc.gatech.edu

Ji Soo Yi
Purdue University
yij@purdue.edu

ABSTRACT

Despite recent improvements in various computational approaches such as machine learning, natural language processing, and computational linguistics, making a computer understand unstructured, human-generated text still remains a difficult problem to solve. To alleviate the challenges, we propose an approach called "Opinion Marks," which enables writers to mark positive and negative aspects of a topic on their own text. In addition, Opinion Marks incorporates an automatic marking suggestion algorithm to offload a user's marking effort. The phrases marked with Opinion Marks can be further used to clarify sentiments of other text in the similar context. We implemented Opinion Marks on a question answering website **http://caniask.net**. To test the efficacy of Opinion Marks, we conducted a crowdsourcing-based experiment with 144 participants in a between-subject design under the three different conditions: 1) human marking only; 2) machine marking only (automatic marking suggestion); and 3) human-machine collaboration (Opinion Marks). This study revealed that Opinion Marks significantly improves the quality of marked phrases and usability of the system.

Keywords

human-based computation; user interface; crowdsourcing

Categories and Subject Descriptors

H.5.m. [**Information Interfaces and Presentation (e.g. HCI)**]: Miscellaneous

General Terms

Human Factors; Design; Experimentation.

1. INTRODUCTION

On the web, user-generated, unstructured text documents (hereinafter called text) are being rapidly generated in various forms such as product reviews (e.g., **http://www.amazon.com**), question answering (e.g., **http://www.ask.com**), and discussion forums (e.g., **http://www.ubuntuforums.org**). As the volume of text grows, it is becoming more challenging for people to efficiently grasp useful information such as others' opinions and recommendations.

Computational approaches developed by natural language processing and machine learning have made notable progresses in delivering a great quality of summaries out of copious text. However, these techniques done in a fully automated manner bear inherent limitations in capturing the true semantic meanings and intentions that writers intend to convey.

Alternatively, beyond fully automated computational approaches, human-computing approaches [27], which leverage human capabilities in computational steps, have been gaining popularity. Many of the semi-supervised learning methods from machine learning areas assume human input as a main source of additional supervision, which often lead to significant improvement in desired tasks.

Nonetheless, the main issue is how to support and encourage users so that they can effortlessly but accurately perform their human-computing tasks (e.g., assigning labels), while doing their original jobs (e.g., writing on the web). In the context of microblogging and social networking services, a representative example is a user-generated hashtag, a word tag with a prefix symbol "#". Through small efforts taken by users, hashtags have been shown useful when other users group and filter microblogs, which would otherwise be difficult [6, 12].

Motivated by these progresses, we propose Opinion Marks, an advanced human-based computing technique that allows users to mark their opinions during writing processes in an efficient, user-friendly manner. Basically, Opinion Marks assumes three different possible components available in textual data representing humans' opinions: (1) a topic to be discussed, (2) a positive aspect, and (3) a negative aspect. Figure 1 shows that a writer can describe a positive aspect ("so sweet") as well as a negative aspect ("too rich") of a particular topic ("Ben & Jerry's") with three surrounding symbols (i.e., **# #**, **+ +**, **- -** for a topic, a positive aspect, and a negative aspect, respectively).

The main goal of Opinion Marks is to provide convenient user interfaces through which users can easily mark the three compo-

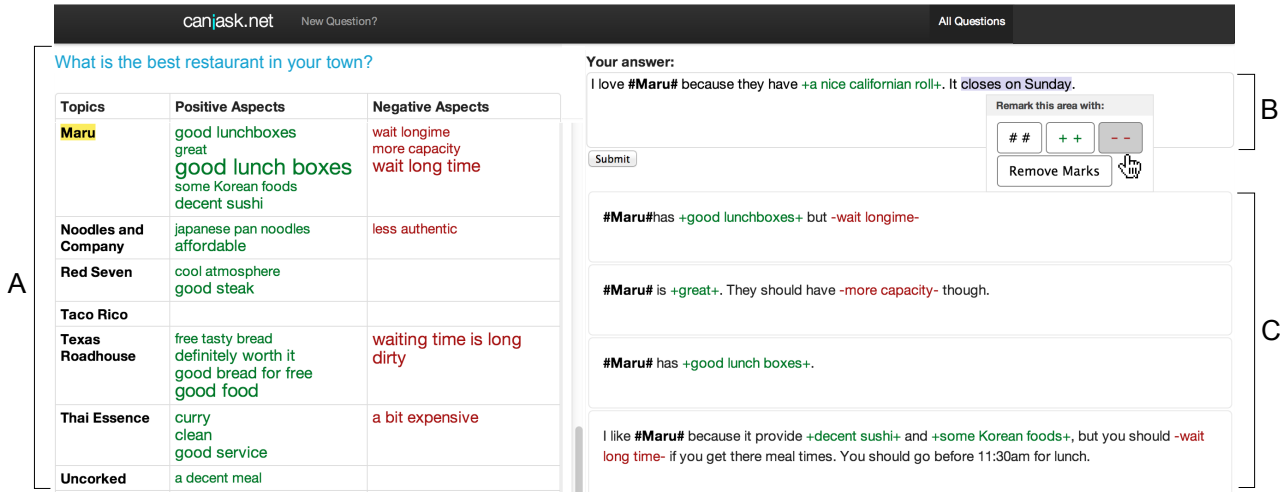


Figure 2: An overview of `http://caniask.net`. (A) Summary Table shows topics, positive aspects, and negative aspects in separate columns from left to right. Users can agree with each positive/negative aspect by clicking each entry, and the font size of each entry reflects the number of agreeing users. In addition, a user can sort columns and filter by topics and aspects, e.g., "Maru" in this case; (B) A user writes his/her own answer and can mark topics, positive aspects, and negative aspects via user interactions provided by Opinion Marks. Upon clicking "Submit," each of the marked phrases is added to Summary Table; (C) Previously created answers are listed. Currently, only those filtered by the topic, "Maru," from (A) were shown.

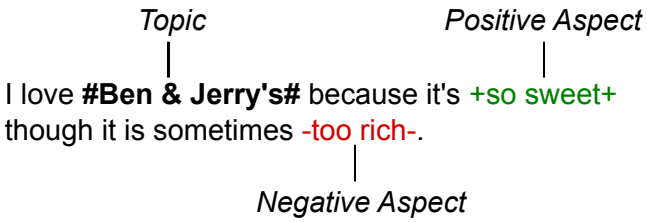


Figure 1: A sentence marked with Opinion Marks. A *topic*, a target entity to be discussed, is surrounded by # # and its *positive aspect* and *negative aspect* by + + and - -.

nents on their text they write in real time. In particular, the novel capabilities of Opinion Marks are as follows:

- As ways to mark topics and their positive/negative aspects during writing phases, we provide a graphical user interface, which is easily accessible by novice users, as well as an inline marking interface, which is geared towards efficient marking processes for experienced users.
- We integrate text mining and natural language processing techniques, such as part-of-speech tagging and lexicon-based sentiment analysis models, to provide recommendations for default marking and encouraging user involvement. Our proposed model adapts itself in response to user corrections.

To deploy Opinion Marks in a familiar environment on the web, we have developed a website, **`http://caniask.net`**, which integrates Opinion Marks in a question answering type of web services (e.g., Quora). Similar to other existing question answering services, our system allows users to freely post questions and write answers to them. When writing an answer, the integrated capabilities of Opinion Marks help users mark topics and positive/negative aspects.

Our paper will present how we substantiated our idea of Opinion Marks in this web-based system and demonstrate its practical efficacy via a crowdsourcing-based user study.

2. OPINION MARKS ON THE WEB

We introduce our publicly available, question answering website, **`http://caniask.net`**, which was used as a testbed for integrating Opinion Marks. We chose to implement a question answering service because we believe that we can induce online discussion evaluating different topics without implementing an full-fledged e-commerce website, which require too much implementation effort. However, we believe that Opinion Marks can be extended to other online services.

The main page of **`http://caniask.net`** (see Figure 2) consists of Textbox with Opinion Marks, Submitted Answers, Summary Table, and Instructions. In **Textbox with Opinion Marks (Figure 2(B))**, users can write actual answers and put markings via user interactions (with a mouse operation or an inline labeling) provided by Opinion Marks. User-marked phrases are highlighted in different colors as shown in Figure 1. Submitted responses are added to the area of **Submitted Answers (Figure 2(C))**, and the parsed results of the submitted response are added to the **Summary Table (Figure 2(A))**. Summary Table provides a comprehensive overview of topics and aspects captured by Opinion Marks from user-generated text. Each row represents a topic along with its collected positive (red-colored) and negative (green-colored) aspects marked by multiple users mentioning the same topic. The font size of positive/negative aspects encodes the number of agreements made by other users (similar to the Like button in Facebook). The Agree button appears when a user hovers around a positive/negative aspect. In addition, clicking on an aspect allows users to filter only the answering posts containing the corresponding keyword as shown in Figure 2(C). To help users better understand how to use Opinion Marks, we provided an instruction when a user fails to mark a topic and an aspect from a sentence. The instruction included two different ways to mark: typing and context menu. We also showed how those marked phrases are inserted into the summary table. All instructions are provided using animated

Graphics Interchange Format (Figure 3 shows selected clips from the instruction).

3. OPINION MARKS

To provide an effective means to instill structures to unstructured online text, we identified two design criteria to maximize the user adoption: First, using this technique should not interrupt the natural process of writing text. Second, using the technique should be intuitive enough for human users to grasp and learn quickly. Opinion Marks achieves these criteria by providing (1) two types of marking interfaces for both experienced as well as inexperienced users and (2) automatic marking suggestion that minimizes human marking efforts as well as encourages user participation in marking tasks. In **`http://caniask.net`**, we implemented Opinion Marks in Textbox. By using Opinion Marks, users can mark topics (what the questioner asked) and their aspects (good and bad things about the topic) on their own answers.

In the following, we describe the details of Opinion Marks in terms of the main concept, user interfaces, and an automatic suggestion module.

3.1 Main Concept

As described in the example in Figure 1, the main concept of Opinion Marks is to mark three types, a topic, its positive aspects, and its negative aspects, from unstructured text that users generate on the web. Such marking is seamlessly integrated into text itself by surrounding particular phrases with special characters # # (topic marks), + + (positive aspect marks), and - - (negative aspect marks). In this sense, Opinion Marks can be considered analogous to a widely-used hashtag, but it conveys more sophisticated information than a hashtag. Considering the great success of a hashtag in the information retrieval context, the marked text via Opinion Marks has significant potential in various text analysis processes such as sentiment analysis and summarization.

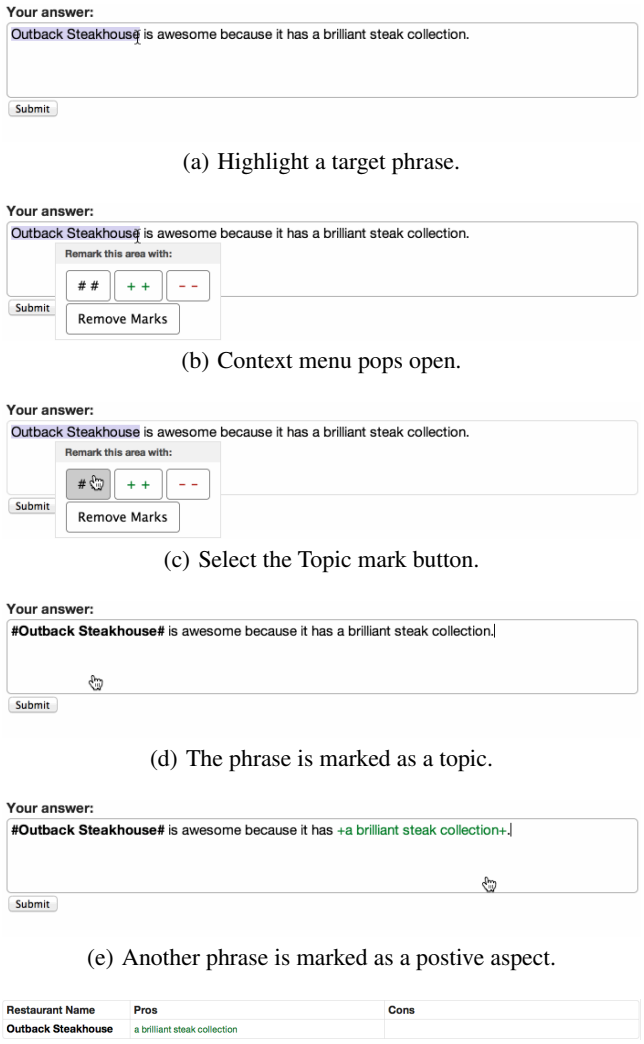
We chose these three special characters, #, +, and -, for Opinion Marks because they are more intuitive than other less frequently used characters (e.g., ^ (caret)). However, the special characters may conflict with other uses of them (e.g., "It's 30+ year old" and "I am a decision-maker"). In order to avoid such conflicts, we detected these characters for Opinion Marks only when they are shown at the beginning of the entire text or preceded by a single white space. For instance, we used the following regular expression to detect the positive aspect surrounded by, + +:

```
/(\s\+{[}\+{]}+\+)|(\+{[}\+{]}+\+)/
```

3.2 User Interface

To support marking processes in a user-friendly and efficient manner, Opinion Marks provide two different user interfaces. The first one is to just let users put the corresponding special characters directly in their text during the writing phase. When a user is familiar with Opinion Marks, this type of a user interface works as the most efficient, straightforward means for marking. In addition, Opinion Marks changes the color of the marked text to bold black (for topics), green (for positive aspects), or red (for negative aspects) on the fly, which is similar to syntax highlighting features in various text editors. In order to further help users, we provided a drop-down menu when a user types the starting mark; the drop-down menu shows topics or aspects entered previously by users who answered the same question.

The second user interface is via a context menu. As Figure 3 shows, once a user highlights a particular phrase via a mouse drag-and-release operation, a custom-designed context menu pops up



(a) Highlight a target phrase.

(b) Context menu pops open.

(c) Select the Topic mark button.

(d) The phrase is marked as a topic.

(e) Another phrase is marked as a postive aspect.

(f) Submission appears on the summary table.

Figure 3: An example of marking target phrases using the context menu in Textbox.

where s/he can select one of the four options: *topic*, *positive*, *negative*, and *remove*. The highlighted phrase will then be marked with the selected option. This user interface provides an inexperienced user with an easy, intuitive interface to select one among the three supported marking types. Furthermore, as will be described in the next section, this type of a user interaction via mouse drag-and-release can also be used to efficiently rectify automatically suggested markings on given text.

3.3 Automatic Marking Suggestion

In addition to the supported user interfaces, we have developed an automatic marking suggestion approach for the parts of text that have to be potentially marked but are not done so yet by users. Given that any single state-of-the-art techniques cannot fully catch all the human semantics and intent, the main purpose of this module is not to generate perfectly accurate markings in a fully automated manner, but deliver to users our best-effort candidates to be marked so that we can mitigate human efforts and encourage user participation in marking processes. In this manner, even if an inexperienced user does not mark his/her text, this module will provide machine-

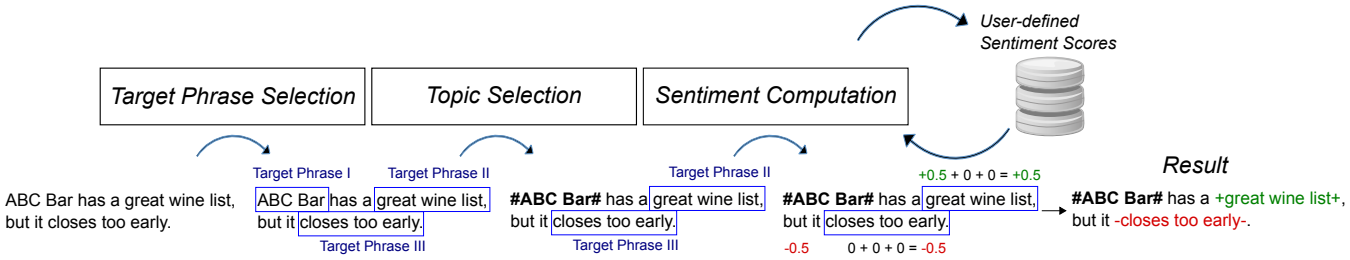


Figure 4: The flow diagram of phrase detection and sentiment computation algorithm.

suggested markings with a reasonable quality, which would make users naturally learn how the system works, revise the suggested markings, and generate their own markings that were missed by the automatic module.

As a first step to create the module, we first deployed and released a minimalistic version of **`http://caniask.net`** in the wild. Then, we investigated how users marked phrases into topics, positive aspects, and negative aspects. Based on our longitudinal study, our automatic marking suggestion module is built upon three steps: (1) topic/aspect candidate generation, (2) topic selection, and (3) aspect selection based on adaptive sentiment computation. Figure 4 shows the overview of steps of how our automatic marking suggestion works. In the following, we describe each step in detail.

Topic/aspect candidate generation. Given a user-generated textual post, e.g., a single tweet, review article, or comment, this step generates a set of candidate phrases for marking. Basically, we want each of our candidate phrases to be the most meaningful but shortest possible phrase that keeps the author's intent intact. From our preliminary user study where we asked participants to mark text using Opinion Marks (with no automatic suggestion), we found that a majority of marked phrases are noun phrases and verb phrases. Thus, we created a rule-based algorithm that can mimic the marking behavior.

In detail, we first parse each sentence and obtain a part-of-speech (POS) tag[1] for each word in it. Then, we find the main verb for the sentence. If this verb is contained in our predefined set of insignificant verbs with any tense, we exclude them from our candidate phrases since they either do not add much meaning (e.g., be, do, and have) or their meanings can be represented via our positive/negative marks (e.g., like, love, hate, and dislike).

Next, we find nouns or noun phrases used as objects. Then, for each of them, we find and include preceding nouns, pronouns, and adjectives because they are likely to add meanings to it. For the same reason, we include adverbs and preposition phrases following each of main nouns. Throughout the process, we achieve a set of candidate phrases $P = \{p_1, p_2, \cdots, p_n\}$, where p_i is in an order of appearance in d.

In addition, we also compare this candidate phrase set with topics and phrases that are marked by other users previously. If we find a more comprehensive and inclusive phrase in previous phrases, then we used the phrase instead. This was done to reflect human-marked phrases from other users.

Topic selection. The next step is to determine a topic phrase p_t from P. An important assumption here is that a user discusses only a single topic in each post d. For instance, in case of an online review, a restaurant review, even though a user could discuss various aspects such as service, food quality, atmosphere, we assume that s/he talks about a particular restaurant, which would be marked as a topic.

Based on this assumption, we select p_t as the first appearing p_i (the smallest i value) such that p_i satisfies either of the two conditions: (1) p_i is tagged as a subject in its corresponding sentence or (2) any noun in p_i starts with a capital letter. When two condition conflicts, we respected (2) because the capitalized nouns show author's more explicit intent. For example, if an author writes "I feel McDonald's is great because they have cheap and nice burgers", then McDonald's will be captured as a topic because this is the first capitalized noun that appears in the sentence.

Aspect selection with adaptive sentiment computation. Now, we select positive or negative aspects from $P \backslash \{p_t\}$. Our basic approach is a lexicon-based sentiment analysis approach that we modified from a previously developed algorithm [22]. That is, we compute an overall sentiment score $S(p_i)$ for p_i by aggregating the word-level sentiment scores for words contained in p_i. Specifically, suppose p_i is composed of a sequence of m_i words, i.e., $p_i = (p_{i,1}, p_{i,2}, \cdots, p_{i,m_i})$ where $p_{i,j}$ represents the j-th word in p_i. Assuming that an word-level sentiment score $s(w_j)$ for a word w_j is defined, $S(p_i)$ is computed as

$$S(p_i) = \sum_{i=1}^{m_i} s(p_{i,j}).$$

After computing $S(p_i)$ for each p_i in $P \backslash \{p_t\}$, we select positive aspects as those p_i's with $S(p_i) \geq \delta_+$ and negative ones for those with $S(p_i) \leq \delta_-$ where we set $\delta_+ = 0.3$ and $\delta_- = -0.3$.

Now, let us describe how we determine $s(w_j)$, which we call a crowd-driven sentiment score. Initially, we start with a set of words each of which, w_j, is assigned a predefined score $\hat{s}(w_j)$ ranging from -1 (negative) to 1 (positive)[2]. Starting with these predefined scores, we adaptively adjust our crowd-driven sentiment score $s(w_j)$ for w_j during user marking processes, i.e.,

$$s(w_j) = \begin{cases} \hat{s}(w_j) & \Delta(w_j) \leq 0 \\ \left(1 - \frac{\Delta(w_j)}{N}\right) \hat{s}(w_j) & 0 < \Delta(w_j) \leq N \\ -\operatorname{sgn}\left(\hat{s}(w_j)\right) \frac{\exp(k(\Delta(w_j)-N))-1}{\exp(k(\Delta(w_j)-N))+1} & \Delta(w_j) > N \end{cases} \quad (1)$$

where k and N are parameters (e.g., $k = 0.1$ and $N_{thres} = 5$ in our case). In addition, $\Delta(w_j)$ is defined as

$$\Delta(w_j) = -\operatorname{sgn}\left(\hat{s}(w_j)\right) \Delta_{p-n}(w_j)$$

where $\Delta_{p-n}(w_j)$ represents the occurrence count of w_j among ex-

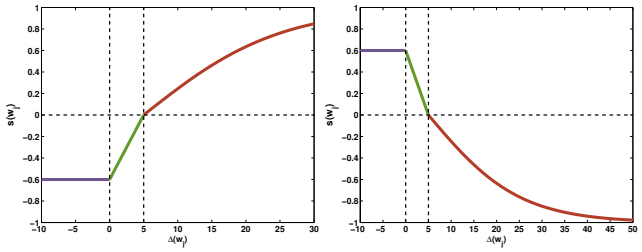
[1]We used Stanford Natural Language Processing library available at **`http://www-nlp.stanford.edu/software/index.shtml`**.

[2]We obtained the word-score list from **`https://github.com/cmaclell/Basic-Tweet-Sentiment-Analyzer`**.

isting positive aspects minus that of w_j among existing negative aspects. Intuitively, the value of $\Delta(w_j)$ represents how often opposite sentiments have been observed relatively to agreeing sentiments with respect to the predefined sentiment polarity of w_j, i.e., $\text{sgn}(\hat{s}(w_j))$. Eq. (1) reflects such observations and adjusts the sentiment $s(w_j)$ accordingly. Figure 5 shows the example functions of $s(w_j)$ depending on $\Delta(w_j)$. In Figure 5(a), starting from an initially negative sentiment score, $\hat{s}(w_j) = -0.6$, as we observe more examples with the opposite sentiment polarity, i.e., increasing $\Delta(w_j)$, $s(w_j)$ changes gradually from a negative value to a positive one. The parameter N, e.g., 5 in our case, determines a particular value of $\Delta(w_j)$ where the original sentiment polarity starts to be reversed. Figure 5(b) shows another example with an initially positive sentiment score. Depending on the size of a text corpus, one can change the parameter values of N and k.

In practice, this measure plays a role of reflecting (1) the context in which the textual post was written as well as (2) the general impression of users associated with a particular facet. That is, in the case of the former, a particular word may have a different sentiment depending on the context. For instance, the word "expensive" can be used with a positive sentiment in a casual conversations, say, in "Wow! you wear an expensive watch!" while it is with a negative sentiment when it comes to most of the product reviews. On the other hand, in the case of the latter, we aim at taking the general positive or negative opinion about a particular facet that takes users into account even for those words initially with no sentiments attached. For example, suppose a previously marked review article about a restaurant is available as "I had to -wait too long on lunch time-." Suppose also that a newly created but yet unmarked review is posted as "You need to get there before 11:30am for lunch." In this example, the word "lunch" was contained in a negative aspect in the first review. Now, in the second one, suppose we need to determine the sentiment on the candidate phrase "need to get there before 11:30am for lunch." In this phrase, there are possibly no words associated with negative sentiment. However, the user-defined sentiment score for the word "lunch" will have a negative value due to the previous marking, which would result in suggesting the candidate phrase as a negative aspect.

Finally, we handle those conjunctions changing the polarity of a sentiment (e.g., "but," "however," and "on the other hand") as follows: If a candidate phrase appears after such a conjunction, we add the opposite sentiment score from the phrase before the conjunction. This sentence-level correction was placed to reflect user's intent more accurately.



(a) $\hat{s}(w_j) = -0.6$ (b) $\hat{s}(w_j) = 0.6$

Figure 5: The example graphs of $s(w_j)$ vs. $\Delta(w_j)$. We used $k = 0.1$ and $N = 5$.

3.4 Method

To evaluate how the automatic marking suggestions influence users' adoption and correct use of marking, we conducted a crowdsourcing based user study. We asked a total of 204 participants to answer the question, "What is your favorite fast-food restaurant?" To analyze the effectiveness of the different features of Opinion Marks, participants were assigned randomly with one of three conditions: 1) 77 participants with Machine marking only (***M-only***); 2) 67 participants with Human marking only (***H-only***); and 3) 60 participants with Human + Machine marking (***H+M***). In all the three conditions, the website first provided a brief instruction on how to use the three mark types (# #, + +, and - -). The three conditions differ in how our system behaves after a participant initially submits an answer. In *M-only*, once a participant submits an answer, it is marked by the automatic marking suggestion module. This automatically marked answer is considered final, so no additional user interaction is allowed after the initial submission. In *H-only*, if a participant submits an answer without any markings on, the website shows an animated instruction explaining how to use markings. Then, the website returns the original unmarked answer to the participant and gives another chance to put markings and resubmit the marked answer. In *H+M*, the procedure is the same as *H-only* except when a participant submits an answer without any markings, the website returns the answer containing automatically marked phrases.

3.5 Results

User Adoption. Table 1 shows the percentage of participants' submissions with marks in their initial and final submissions, regardless of whether their markings were properly done or not. As expected, few participants adopted Opinion Marks in their initial submissions. About 1 or 2 out of 10 participants used marking in their initial submissions (i.e., 14.29% for *M-only*, 10.45% for *H-only*, and 22.64% for *H+M*), which are not significantly different from each other ($\chi^2(2, N = 204) = 2.299, p = 0.313$). After initial submission, in *H-only*, despite the additional instruction on how to mark, 26 (38.81%) participants still did not adopt markings. In *H+M*, 41 (68.33%) participants refined their markings after seeing the suggested markings, which shows statistically significant difference ($\chi^2(1, N = 108) = 33.191, p < 0.001$). This is interesting because 41 out of 47 participants, who received answers with default marks from automatic suggestion, did not keep default markings even though it is easier to do so. They rather refined their markings to show their clear intent. This shows that automatic marking suggestions clearly nudge participants to adopt markings.

Table 1: Number of participants submitted marked/unmarked answers.

Initial Submission Final Submission	Marked -	Unmarked Unmarked	Unmarked Marked
M-only	11 (14.29%)	-[a]	66 (85.71%)
H-only	7 (10.45%)	26 (38.81%)	34 (50.75%)
H+M	12 (20.00%)	-[a]	48 (80.00%)[b]

[a] In *M-only* and *H+M*, since each submission is marked by automatic marking suggestion, there is no unmarked cases in their final submissions.

[b] Out of 48 participants, 7 participants kept the machine suggested markings, but the rest of 41 manually refined the machine-suggested markings.

Marking Correctness. To evaluate the correctness of markings, two of the authors codified each submission into seven error types.[3] First, reversed sentiments (RS) indicate that aspect phrases are marked with an opposite sentiment, e.g., "+many people on weekend rush+", which should have been marked with "--". Second, incorrect phrase types (IP) indicate that topic phrases

[3]To avoid any potential biases, we first removed the information in which conditions each answer was written.

are marked as aspect phrases or vice versa, e.g., "**#delicious biscuit#**". Third, fragmented phrase (FP) indicates that a single markable phrase is cut into multiple phrases with marks, e.g., "+lots of +offers+ and +discounts+ on +festival days+". Fourth, merged phrase (MP) indicates that a marked phrase includes unnecessary words so it should have been shorter or should have been broken into multiple phrases, e.g., "-It has a lot of waiting issue and need to do advance booking so it appears as a con to me-" and "+Fresh ingredients and a lot of choices+". Fifth, unmatched marks (UM) indicate that phrases are marked with different starting and ending marks, e.g., "#Five Guys+". Sixth, missed markable phrase (MM) indicates that the submission has markable phrases that are not marked at all. Seventh, falsely marked phrase (FM) indicates that the submission includes phrases that should not have been marked but were marked.

To test the effects of the three conditions on each error type, we conducted logistic regression analysis with Type III tests to fit the result and computed the odds ratios with 95% confidence intervals. The results show that *H+M* generate less errors in two error types (MM and FP) that *H-only* and *M-only* are likely to generate. Participants in *H+M* generated less errors than those in *H-only* for missing marks on markable phrases (MM, $F(2,201) = 68.12$, $p < 0.001$, Figure 6 (f)); participants in *H+M* generated less errors than those in *M-only* for breaking a single phrase with several marks (FP, $F(2,201) = 10.12$, $p = 0.006$, Figure 6 (g)). Furthermore, participants in *H+M* did not show significantly higher or less errors for the rest of the error types. In contrast, participants in *H-only* made more mistakes than *M-only* in properly enclosing phrases with starting and ending marks (UM, $F(2,201) = 6.87$, $p = 0.032$, Figure 6 (e)) and marking all of markable phrases (MM, $F(2,201) = 68.12$, $p < 0.001$, Figure 6 (f)). On the other hand, participants in *M-only* made more errors than *H-only* in marking non-markable phrases (FM, $F(2,201) = 6.06$, $p = 0.048$, Figure 6 (c)) or breaking a single markable phrase into several phrases (FP, $F(2,201) = 10.12$, $p = 0.006$, Figure 6 (g)) than those in *H-only*. The distinctive error distributions show limitations of *H-* and *M-only*.

The experiment results show that automatic marking suggestion increase the correct adoption of Opinion Marks. We had a concern over implementing automatic suggestions: people may not revise the default marks, thereby leading to generating unsupervised, incorrect phrases. This was wrong–people revised phrases after suggestion, so we could collect more phrases. This iterative loop between machine suggestion and human correction makes positive impacts on the output quality. The result shows that users can "do more and better" with "a slight nudge".

We see interesting implications about human-machine collaboration on the Opinion Marks tasks in Figure 6. Automatic suggestion cannot mark phrases accurately so that they deliver users' intent. It makes mistakes in falsely marking phrases that are not markable (see Figure 6 (g)) and in separating phrases that should be merged to one (see Figure 6 (c)). On the other hand, humans make errors in using the markups accurately (see Figure 6 (e)) and in putting marks on phrases that should be marked (see Figure 6 (f)). Our suggestion, tightly integrating *H+M*, indeed improved some weaknesses on both sides and result in the higher marking rates in the submitted answers.

4. DISCUSSION

As seen in the previous section, we demonstrated that Opinion Marks allows us to collect accurate and compact phrases along with their sentiments by leveraging both machine learning and human computation. In this section, we discuss the potential impact of Opinion Marks from two different perspectives: (1) faithful support for machine learning and (2) broad applicability to real-world domains.

First, the main idea of Opinion Marks, which converts unstructured text into a structured form, can potentially boost the performance of various machine learning techniques in text analysis. Previously, the first step in this domain is usually to preprocess unstructured text data using a bag-of-words encoding scheme, which basically uses all the available keywords in a corpus regardless of their respective importance and noise. On the other hand, the phrases collected via Opinion Marks provide a higher quality of data representations with a selective set of meaningful keywords. Furthermore, these keywords are represented at a right level of granularity based on human understanding, which is neither a word level (fine-grained) nor a document level (coarse-grained). In this sense, our work can trigger significant improvement in machine learning tasks, such as topic modeling, document summarization, and sentiment analysis.

Second, although we currently applied Opinion Marks to our own proof-of-concept system, **`http://caniask.net`**, it can be easily embedded into broader online writing environments, such as product reviews, social media, and discussion forums. The user interfaces that we designed for Opinion Marks would not require major modifications for its integration to an existing system. Upon integration, Opinion Marks could be utilized in efficiently marking massive-scale data already available on the system as well as rapidly generated data, e.g., Amazon product reviews. In this manner, Opinion Marks will bring tremendous impact by accelerating a tedious process of reading individual text much fast. Alternatively, one could create a pluggable online writing platform equipped with Opinion Marks. For example, a web-based service available at **`http://disqus.com`** provides users with a discussion thread platform that can be integrated into their own websites. In this manner, the impact of Opinion Marks can quickly reach various real-world domains.

Third, we believe that our technique can be applied to collect accurate phrases from existing large-scale online text (e.g., product reviews) written by other users. One may have a concern that our technique may not work because users may not want to adopt this technique while writing their text due to the lack of explicit motivation. In our experiment, however, we showed that crowdsourced workers can successfully mark accurate phrases from their own text. The difficulty of task (i.e., marking opinion phrases from text) will be equal for text written by other users. Thus, we claim that our technique can be flexibly applied either by writers or by analysts (e.g., crowdsourced workers) to create a human-readable summary out of copious text. Of course, our "agreement" measures in summary table are in place to motivate users by social-psychological incentive. In addition to such monetary and social incentives, we can also utilize the gamification idea for the phrase extraction process.

5. RELATED WORK

In this section, we discuss related work from two perspectives: computational approaches (e.g., sentiment analysis and phrase mining) and human-computing approaches.

5.1 Computational Approaches

A myriad of automated computational approaches for text analysis have been proposed from machine learning, data mining, information retrieval, natural language processing, and computational linguistics. In general, most of them aim at revealing meaningful insights and summaries out of unstructured text data. In the follow-

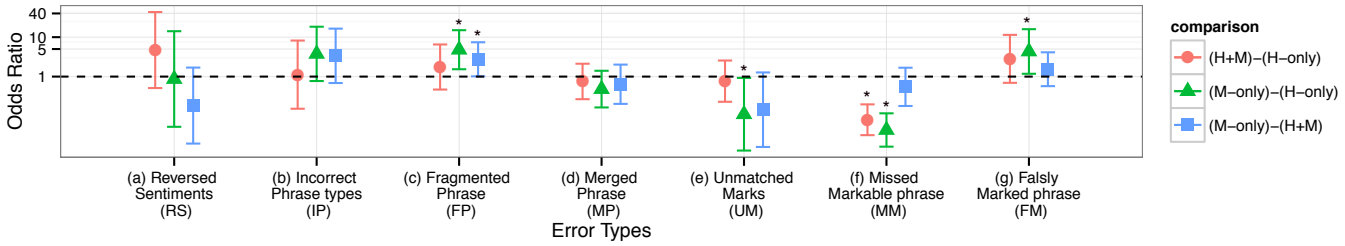


Figure 6: The 95% confidence intervals of odds ratios of seven error types in pairwise comparison between three experimental conditions. The asterisk mark (*) indicates statistically significant odds ratios.

ing, we review two closely related areas to our work: (1) sentiment analysis/opinion mining and (2) key phrase extraction.

Sentiment analysis and opinion mining [16, 21] intend to detect contextual polarity of given text information, e.g., a positive or a negative sentiment. Different methods work at a different level, such as a keyword, a phrase, a sentence, or an entire document [19, 24, 31]. Traditionally, a simple lexicon-based approach, which aggregates word-level sentiment scores, has been widely used [10], and until recently, numerous advanced methods have also tried to capture subtle connotations in human language, context dependency, and incorrectness [3, 5, 26, 32]. Beyond traditional online review data, sentiment analysis has been actively applied to novel social media data [2, 18]. Nonetheless, sentiment analysis is still an active area of research, which shows the significant difficulty of the problem. In addition, in many approaches, it is not usually a main concern to extract the key phrases that directly support detected sentiments out of an entire text.

On the other hand, various approaches for key phrase mining have also been proposed [8, 34]. In a sense that these key phrases are useful in summarizing documents and revealing high-level topics, it has often been studied together with topic modeling [14, 17, 30]. However, most of these methods rely on the frequent occurrence of a particular phrase in the exact same form, and thus they cannot properly detect different phrases with the same meaning [25].

Generally, perfectly understanding writers' meaning and intent is still a fundamentally challenging task when only using fully automated approaches.

5.2 Visualization Approaches

Visualization approaches, often used along with other computational approaches, provide users with a gateway to interactively explore text data. Many developments have been made particularly in context of online reviews. OpinionBlocks provide the overview of snippets from multiple consumer reviews [1]. Oelke et al. [20] presented features and sentiments in a matrix visualization. Review Spotlight shows word clouds of useful adjective-noun word pairs [33]. Similarly, ReCloud also provides word cloud of online reviews [29]. RevMiner and Odin provide mobile interfaces for users to explore reviews based on opinion mining [9, 11]. Termite uses a matrix visualization to show coocurrences of key topic words appearing in documents [4]. These approaches rely upon computation approaches to extract features (e.g., sentiment) to visualize. Therefore, it is necessary to improve the phrase and sentiment extraction tasks for visualizations to be unbiased and useful.

5.3 Human-Computing Approaches

Human-computing approaches aim to delegate part of jobs for machines to human workers, so that they can achieve the best outcome [7, 27]. Such human computing approaches have been successfully applied in image tagging [28], image segmentation [23], and text categorization tasks [13], which lead to better performances of diverse machine learning algorithms.

In the context of sentiment analysis and opinion mining, *Opinion Observer* provides an interface for user correction on analyzed sentiment results from product review data [15]. More recently, a web-based sentiment analysis system where a user can run sentiment analysis on his/her own text and make corrections has been proposed [24]. However, as far as our knowledge goes, none of the previous systems have smoothly integrated inline labeling approaches to the writing phrase while preserving a user's semantic context. Such integration is crucial in maximizing user participation and accuracy in the human labeling processes [28].

6. CONCLUSIONS

In this paper, we presented Opinion Marks and its proof-of-concept system, **`http://caniask.net`**, which enables users to mark their opinions while they write text. To this end, we designed two interaction methods to add such marks with minimal human efforts. Opinion Marks also features automatic marking suggestion based on our carefully designed algorithm so that we can maximize user participation as well as accuracy. Our crowdsourcing-based user study demonstrates that Opinion Marks successfully leveraged the collaborative effects between human users and computer machines for enhancing the output quality and user participation.

Our work has great potential in diverse online writing applications. Specifically, as our future work, we plan to analyze large-scale document data such as Amazon.com product reviews based on Opinion Marks. In this scenario, we will investigate how to further improve the efficiency of a marking process given numerous text data by using more advanced automatic suggestion algorithms as well as more convenient user interfaces. In addition, instead of just a tabular-style display, we will focus on how to effectively summarize the marked entries and support humans for better understanding of them.

7. REFERENCES

[1] B. Alper, H. Yang, E. Haber, and E. Kandogan. OpinionBlocks: visualizing consumer reviews. In *Interactive Visual Text Analytics for Decision Making in conjunction with VisWeek 2011*, Providence, RI, 2011.

[2] L. Chen, W. Wang, M. Nagarajan, S. Wang, and A. P. Sheth. Extracting diverse sentiment expressions with target-dependent polarity from twitter. In *Proceedings the 6th International AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 50–57, 2012.

[3] Y. Choi and C. Cardie. Learning with compositional semantics as structural inference for subsentential sentiment analysis. In *Proceedings of the Conference on Empirical*

Methods in Natural Language Processing, pages 793–801, 2008.

[4] J. Chuang, C. D. Manning, and J. Heer. Termite: Visualization techniques for assessing textual topic models. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, AVI '12, pages 74âĂŞ–77, New York, NY, USA, 2012. ACM.

[5] X. Ding, B. Liu, and P. S. Yu. A holistic lexicon-based approach to opinion mining. In *Proceedings the International Conference on Web Search and Data Mining (WSDM)*, pages 231–240, 2008.

[6] M. Efron. Hashtag retrieval in a microblogging environment. In *Proceedings the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 787–788, 2010.

[7] S. Farnham, H. R. Chesley, D. E. McGhee, R. Kawal, and J. Landau. Structured online interactions: improving the decision-making of small discussion groups. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*, pages 299–308, 2000.

[8] D. Fetterly, M. Manasse, and M. Najork. Detecting phrase-level duplication on the world wide web. In *Proceedings the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 170–177, 2005.

[9] J. M. Hailpern and B. A. Huberman. Odin: Contextual document opinions on the go. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pages 1525–âĂŞ1534, New York, NY, USA, 2014. ACM.

[10] M. Hu and B. Liu. Mining and summarizing customer reviews. In *Proceedings of the 10th ACM International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 168–177, 2004.

[11] J. Huang, O. Etzioni, L. Zettlemoyer, K. Clark, and C. Lee. RevMiner: an extractive interface for navigating reviews on a smartphone. In *Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology*, UIST '12, pages 3–âĂŞ12, New York, NY, USA, 2012. ACM.

[12] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proceedings the 19th International Conference on World Wide Web (WWW)*, pages 591–600, 2010.

[13] A. C. KÃűnig and E. Brill. Reducing the human overhead in text categorization. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 598–603, 2006.

[14] B. Liu, C. W. Chin, and H. T. Ng. Mining topic-specific concepts and definitions on the web. In *Proceedings the 12th International Conference on World Wide Web (WWW)*, pages 251–260, 2003.

[15] B. Liu, M. Hu, and J. Cheng. Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of the 14th international conference on World Wide Web*, pages 342–351, 2005.

[16] B. Liu and L. Zhang. A survey of opinion mining and sentiment analysis. In *Mining Text Data*, pages 415–463. Springer, 2012.

[17] Z. Liu, W. Huang, Y. Zheng, and M. Sun. Automatic keyphrase extraction via topic decomposition. In *Proceedings the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 366–376, 2010.

[18] Y. Mejova and P. Srinivasan. Crossing media streams with sentiment: Domain adaptation in blogs, reviews and twitter. In *Proceedings the 6th International AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 234–241, 2012.

[19] T. Nakagawa, K. Inui, and S. Kurohashi. Dependency tree-based sentiment classification using crfs with hidden variables. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 786–794, 2010.

[20] D. Oelke, M. Hao, C. Rohrdantz, D. A. Keim, U. Dayal, L. E. Haug, and H. Janetzko. Visual opinion analysis of customer feedback data. In *IEEE Symposium on Visual Analytics Science and Technology, 2009. VAST 2009*, pages 187–194. IEEE, 2009.

[21] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2008.

[22] C. Rohrdantz, M. C. Hao, U. Dayal, L.-E. Haug, and D. A. Keim. Feature-based visual sentiment analysis of text document streams. *ACM Transaction of Intelligent Systems and Technology.*, 3(2):26:1–26:25, Feb. 2012.

[23] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. Labelme: A database and web-based tool for image annotation. *International Journal of Computer Vision*, 77(1-3):157–173, May 2008.

[24] R. Socher, A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Conference on Empirical Methods in Natural Language Processing*, 2013.

[25] V. Stoyanov, N. Gilbert, C. Cardie, and E. Riloff. Conundrums in noun phrase coreference resolution: Making sense of the state-of-the-art. In *Proceedings the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*, pages 656–664, 2009.

[26] R. Trivedi and J. Eisenstein. Discourse connectors for latent subjectivity in sentiment analysis. In *Proceedings the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 808–813, 2013.

[27] L. von Ahn. Human computation. In *Proceedings of the 46th Annual Design Automation Conference*, pages 418–419, 2009.

[28] L. von Ahn and L. Dabbish. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 319–326, 2004.

[29] J. Wang, J. Zhao, S. Guo, C. North, and N. Ramakrishnan. ReCloud: semantics-based word cloud visualization of user reviews. In *Proceedings of the 2014 Graphics Interface Conference*, pages 151âĂŞ–158, Toronto, Ont., Canada, Canada, 2014. Canadian Information Processing Society.

[30] X. Wang, A. McCallum, and X. Wei. Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In *Proceedings the 7th IEEE International Conference on Data Mining (ICDM)*, pages 697–702, 2007.

[31] T. Wilson, J. Wiebe, and P. 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, 2005.

[32] Y. Wu, Q. Zhang, X. Huang, and L. Wu. Phrase dependency parsing for opinion mining. In *Prof. the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1533–1541, 2009.

[33] K. Yatani, M. Novati, A. Trusty, and K. N. Truong. Review spotlight: a user interface for summarizing user-generated reviews using adjective-noun word pairs. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, pages 1541–1550, New York, NY, USA, 2011. ACM.

[34] H. Zha. Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering. In *Proceedings the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 113–120, 2002.