Identifying Pros and Cons of Product Aspects Based on Customer Reviews

Ebad Ahmadzadeh
Dept. of Computer Sciences
Florida Institute of Technology
Melbourne, Florida 32901
Email: mahmadzadehe2012@my.fit.edu

Philip K. Chan
Dept. of Computer Sciences
Florida Institute of Technology
Melbourne, Florida 32901
Email: pkc@cs.fit.edu

***Abstract*—The task of identifying pros and cons from product reviews has applications in decision support for consumers. It becomes even more useful when the pros and cons are identified for product aspects so consumers can quickly see strengths and weaknesses of each aspect of the product without reading all reviews. Given a collection of product reviews, we automatically extract relevant product aspects, find the most significant sentences that represent pros and cons for each aspect, and provide a summary for each aspect. We introduce SS2 to select sentences that are likely to represent pros/cons and are semantically related to the aspect to which they are associated. Our results on three data sets indicate that compared to an existing algorithm, our algorithm can generate more meaningful summarized aspects, along with a list of pros and cons more closely related to each aspect.**

I. Introduction

In the recent years, the volume of user-generated social media content has been growing increasingly. People across the Web are constantly sharing experiences and opinions about a wide range of situations. Many research lines have focused on using this information as a data source to apply to different domains such as decision support, question answering, machine translation, etc. Reviews of products and services is one of the highly valuable sources of data that can be used to answer interesting questions about the product or service. For example, identifying strengths and weaknesses of a product can be useful for the company that makes the product to improve the weaknesses and add desired features.

In this work we propose an algorithm that extracts product aspects (e.g. camera can be an aspect when the product is a cellphone.) and identifies weaknesses and strengths for each aspect in the form of pros and cons. We also provide a summary for each product aspect so readers can understand each extracted aspect in a glance before reading the selected pro and con sentences.

Fig. 1 shows the current presentation of Amazon reviews that displays top positive and negative reviews as well as the distribution of ratings on a scale of one to five. The current presentation of the reviews has two main shortcomings; First, it is not convenient for users to find reviews about a certain topic or aspect of the product because they are organized in a linear list. Second, consumers consider different aspects of the product to make the buying decision. But it can be quite time consuming for them to find out others' opinion about the aspects that matters to them. Fig. 2 illustrates an example of a user interface based on our algorithms that could be shown to users to further assist them in making the buying decision. In this case the product is a coffee maker. The product aspects (e.g. cup size, coffee quality, water reservoir, etc.) are extracted by our algorithm, and users can expand each aspect to view the pros and cons related to the aspect based on the available reviews. We organize the reviews under product aspects. So, the consumers can quickly find the reviews related to the aspects. Moreover, for each aspect we provide pros and cons. Therefore, the consumers can quickly observe the opinion of others about each aspect.

Our contributions include:

1) jointly identifying product aspects and pros and cons with respect to each aspect,
2) summarizing the aspects in the form of bigrams that show different descriptions or opinions about each product aspect
3) proposing modified Significant Score (SS2) with additional factors to quantify significance of sentences in terms of representing meaningful pros or cons with respect to the product aspects, and
4) based on three data sets from Amazon reviews, showing that our algorithm finds pros and cons that are more meaningful and related to the aspects presented in a summarized form compared to [19]

We discuss the related work in Sec. II. Sec. III provides the problem statement and describes the different steps of our algorithm. We evaluate our algorithms in Sec. V and conclude in Sec. VI.

II. Related Work

The task of identifying pros and cons from product reviews usually involves three procedures; identifying product aspects, sentiment analysis, and summarization. We briefly review some existing algorithms for each task as they relate to our task.

The most popular techniques for identifying product aspects are based on frequency analysis [1]–[3]. Zhao et al. [4] propose a technique based on syntactic structures. Also other

studies have explored supervised [5] and unsupervised [6] techniques. Aspect-based sentiment analysis can be done as a joint task where the goal is to calculate sentiment score for each product aspect [7], [8].

Sentiment analysis can be done at different levels such as word, phrase, sentence and document. VADER [9] is an unsupervised, lexicon and rule-based method tuned for sentiments of words, phrases and sentences expressed in social media. Among the state of the art supervised techniques those of Socher et al. [10] can be mentioned. Moreover, document-level sentiment classification techniques, often applied to reviews, have been explored extensively [11], [12]. The task of distinguishing between subjective and objective expressions is useful to separate opinions from facts [13], [14] with applications in question answering, summarization, etc. Subjective clues were collected as part of the work reported in [15].

The main goal of summarization task is to generate a short but meaningful representation given a larger text. Hu and Lie [1] use feature-opinion pairs to summarize reviews. Lu et al. [16] have proposed techniques based on clustering of phrases and aspects.

As discussed above, many of the existing research focus on one or two of the tasks but not all three. For instance, Zhao et al. [4] focus on aspect extraction, Hai et al. [17] propose a supervised technique for joint modeling of aspects and sentiment, but they do not provide a solution to summarize the reviews in the form of pros and cons. Ahmadzadeh and Chan [18] propose a method to identify pros and cons of doing actions based on social media. However, they do not extract aspects. Also, their solution is based on events before and after doing the action (e.g. purchasing the product in this case). But product reviews usually do not include much information about user experience before purchasing the product.

To the best of our knowledge the closest work to ours is that of Kim and Hovy [19], a supervised method for identifying pros and cons from product reviews. They use three categories of features to train; 1) Lexical Features consist of unigrams, bigrams and trigrams that represent reasoning tokens, 2) Positional Features specify whether the sentence is from the first, second, last or second to the last sentence of the review paragraph, 3) Opinion-bearing Word Features consist of a dictionary of pre-selected opinion words. Each learning instance is a sentence associated with a label ("pro", "con" or "neither"). They separate the task of finding pro and con sentences into two phases each being a binary classification. In the first phase (identification), they separate "pro" and "con" sentences from "neither", and in the second phase (classification), they classify the candidates into pros and cons. Still they do not provide a method to extract aspects. Therefore, the pros and cons identified by their algorithm is at the product-level. Whereas, in our work we also identify pros and cons at the product aspects.

Figure 1. The current presentation of Amazon reviews

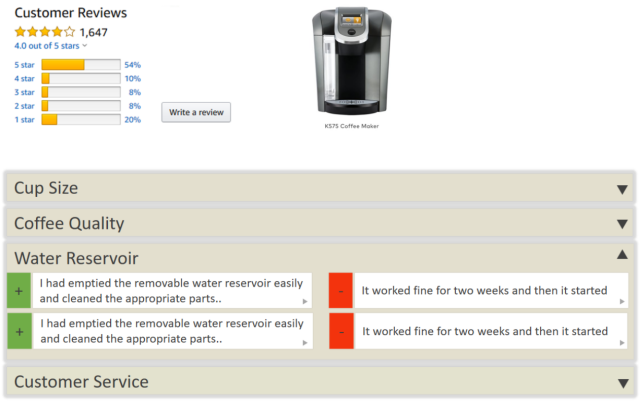

Figure 2. Pros and Cons of a coffee maker product for each extracted product aspect by our algorithms: user is presented by a list of aspects extracted by the algorithm. Clicking on each one expands the list of pros and cons for the aspect based on the reviews.

III. Problem Statement and Approach Overview

Given a collection of product reviews, the goal is to automatically extract relevant product aspects and to find the most significant sentences that represent pros and cons for each aspect. The input is a corpus of product reviews, and the output is M product aspects that are often discussed in the reviews, as well as a list of top K pros and cons for each of the aspects. For example if the product is a cell phone, then aspects could be call quality, price, camera, etc.

Fig. 3 illustrates the overall steps. At the first step, aspects are extracted. Next given the aspect words and aspect reviews, pros and cons are identified. Finally, aspects are summarized and presented by representative bigrams.

IV. Approach

In a more concrete manner as an algorithm, our approach has three main steps as shown in Alg. 1. After preprocessing the review text (e.g. removing URLs, emojis and bad characters) and tokenizing them into sentences, the first step is to find the best set of product aspects and assign the review sentences based on the most probable aspects. Second, we employ a scoring method to select messages that are likely to represent pros or cons. We use this score to identify top K sentences for each topic. Then, we use sentiment intensify score with a minimum threshold to separate pros from cons. Third, we summarize product aspects from each topic and

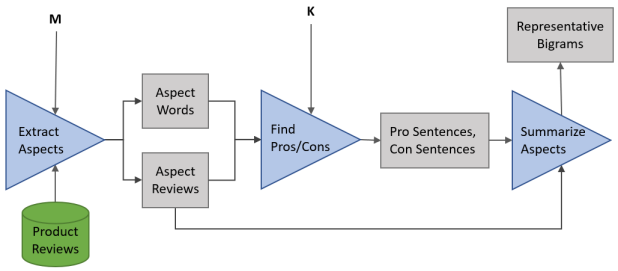

Figure 3. Overall System

present them as bigram phrases using ordered Augmented Expected Mutual Information (AEMI) [20]. We explain each step in more details as follows.

Algorithm 1 mainProCon

Require: $reviews, K, M$

//1: Extract words from aspects and reviews associated with the aspects:
asptWds, asptRev = extractAspects($reviews$, M)
//2: Find top K pros and cons sentences for each aspect:
proSents, conSents = findProsCons(asptWds, asptRev, K)
//3: Summarize aspects in phrases:
aspPhrases = summarizeAspects(asptWds, proSents, conSents)
return proSents, conSents, aspPhrases

A. Aspect Extraction via Topic Modeling

We use LDA [21] to find topics for two main purposes; First, each topic found by LDA is considered to represent an aspect, where each document is represented by a product review. As a result, two tables are generated by LDA; First, aspect-word where the rows are aspects and the columns are words. Each row contains the probability distribution of words for the corresponding aspect. Second, review-aspect where the rows are reviews and columns are aspects. Each row contains the probability distribution of aspects for the corresponding review. The aspect-word table is used as the source to extract aspect words and aspect bigrams, and the review-aspect table is used to assign reviews to the most probable aspects. Therefore, we expect reviews related to each aspect are in the same group.

One of the challenges is to find the best number of aspects for any review corpus as it is an input to LDA algorithm. We use a simple optimization method to identify the best number of topics (aspects), where the objective is to minimize the overlap between the aspects found by LDA. The overlap is calculated via a weighted intersection of the first M aspect words of each aspect. The weights are those generated by LDA to show the importance of aspect words within each aspect. We run LDA multiple times by increasing the number of aspects. For each run, we calculate the overlap between each pair of aspects.

After finding the best number of aspects, we calculate the aspect-word and review-aspect matrices. We find the top M important words for each aspect using the aspect-word matrix weights generated by LDA. Moreover, we assign each review to the most probable aspect according to the review-aspect matrix. Alg. 2 illustrates the aspect extraction steps.

Algorithm 2 extractAspects

Require: $reviews, M$

//1: Find the best number of aspects by minimizing aspect overlap:
numAspects = minimizeAspectOverlap($reviews$)
//2: Find aspects with LDA:
reviewByAspect, aspectByWord = LDA(reviews, numAspects)
//3: Get the top M words w.r.t the weights in aspectWords matrix:
aspectWords = getTopWords(aspectByWord, M)
//4: Assign each review to the most probable aspect according to revAspects
aspectReviews = assignReviews(reviewByAspect)
return aspectWords, aspectReviews

B. Identifying Pros and Cons

The task of identifying pros and cons has four main steps as shown in Alg. 3; First, quantifying co-occurrence via AEMI [20] between aspect words and other words in order to find the representative word pairs involving the aspects. Second, we select sentences that contain at least one representative word pair with positive correlation. Third, we use a scoring method to rank the sentences within each aspect such that those with higher ranks are more likely to be pros or cons. Fourth, we categorize the ranked sentences into pros and cons. Next we explain each step in more details.

1) Finding Representative Word Pairs for each Sentence: We propose to find representative word pairs such that each pair includes an aspect word and an arbitrary word. Word pairs tend to be more informative than single aspect words because they add a description or opinion about the aspect word. For example, "water reservoir" explains the function of the aspect word "reservoir", and "noisy reservoir" specifies an opinion about the aspect word "reservoir". That is, such word pairs with high co-occurrence are likely representatives of the aspect. Also, sentences that contain such representative word pairs tend to be more significant in terms of representing pros/cons of the product aspects.

We calculate AEMI for all possible pairs of (*word*, *aspectWord*) in the sentences associated to each aspect. *words* can be any word inside the vocabulary constructed from all sentences of an aspect, and *aspectWords* are those from aspect-word table generated by LDA. Equation 1 shows

AEMI for two words in each pair represented by A and B. Each sentence represents an event. $p(A, B)$ is the probability that a sentence contains both A and B, and $p(A, \bar{B})$ is the probability that a sentence contains A but does not contain B. $p(A)$ represents the probability that a sentence contains A, and $p(\bar{A})$ represents the probability that a sentence does not contain A.

Each term of the equation represents the mutual information (MI) for the given pair. The second and the third terms are augmented (A) to Pointwise Mutual Information (PMI). AEMI subtracts the mutual information for events when one member of the pair occurs without the other member. Each term is also weighted by the probability of the pair. That is the summation of the three terms yields the expected (E) mutual information of the pair.

$$\begin{aligned} AEMI(A,B) = p(A,B)log(\frac{p(A,B)}{p(A)p(B)}) \\ -p(A,\bar{B})log(\frac{p(A,\bar{B})}{p(A)p(\bar{B})}) \\ -p(\bar{A},B)log(\frac{p(\bar{A},B)}{p(\bar{A})p(B)}) \end{aligned} \quad (1)$$

We organize them into a table called AEMI table. Each row of the table is a *word* and each column is an *aspectWord* and each cell contains the AEMI value of the corresponding *word* and *aspectWord*. We note that this table is half-filled because the ordering of the words in a pair does not matter. After calculating the AEMI table, we use it to select sentences that contain word pairs with positive AEMI values because those sentences are much more likely to represent pros/cons related to the product aspects. This way, we exclude sentences that do not contain any representative word pairs. Moreover, we identify the best representative word pair (one with highest AEMI) for each sentence. Such representative word pairs associated to each sentence will be used to calculate significance score in Sec. IV-B2 and summarizing aspects in Sec. IV-C.

2) Finding Significant Sentences: We employ Significance Score (SS), a scoring method proposed in [18] and improve it by adding additional factors that can be calculated based on reviews data. By using the new scoring method (SS2) we expect sentences with high scores to have two main characteristics: 1) To find sentences that are likely to represent pros or cons, we use Reasoning, Comparison, Sentiment and Length from the existing factors and add Rating and AEMI as factors. 2) To find sentences that are closely related to the aspect, we introduce Coherence and Coverage as factors.

Reasoning: Sentences with reasoning represent pros and cons of higher quality because they provide reasons for the user opinion. Reasoning factor is calculated as a binary variable that is 1 when any phrase indicating reasoning is observed in the sentence (e.g. because, therefore, etc.), and it is 0 otherwise.

Comparison: Comparison is often used in expressing pros and cons of a product. Comparison factor is also calculated as a binary variable that is 1 when comparison tokens are observed and 0 otherwise. The tokens include both keywords and part-of-speech (POS) tags. We use POS tags that represent comparison words (JJR, RBR, JJS, RBS) [22] as well as a small set of keywords (e.g. more, most, less, enough) to reduce the error due to conversational text.

Length: Number of words in a message is another indication for a message to be informative. This factor is normalized between 0 and 1 by comparing all sentences within a cluster.

Sentiment: We use VADER [9], a rule-based sentiment model to calculate sentiment factor as an aggregated score normalized between -1 and 1 to show negative to positive sentiment respectively.

Coherence: We add Coherence factor to reward selection of sentences that contain word pairs and tend to co-occur closer within the sentence scope, as such pairs are more likely to be semantically related. Given a (keyword, topical word) pair and the corresponding sentence, coherence score is the distance between the position of the topical word and the keyword in the sentence, normalized by the sentence length.

Coverage: Measures the portion of the aspect words that are covered by a given sentence. Therefore, it assigns higher score to sentences that cover more aspect words. We use a weighting mechanism to account for importance of topical words based on the weights in topic-word matrix generated by LDA. Coverage is calculated as sum of the weights of the topical words present in the given sentence, divided by sum of weights of all topical words for the topic.

Rating: Measures the intensity of the review about the product based on the author's opinion. This is very helpful information as very high and low rating values can indicate high likelihood of positive and negative expressions about the product respectively. The review ratings in our datasets are integers between 1 (negative) and 5 (positive). Thus, we use the following formula to emphasize on very high and low values $score = (rating - 3)^2/4$. By subtracting 3 (the middle point in [1, 5]) from $rating$ we calculate how far the rating is from neutral. The result is divided by 4 to normalize the score between 0 and 1. Notice that rating values 1 and 5 both produce the highest value (1) which is desired.

Relevance: Measures the semantic relevance between the aspect words and each sentence. This factor goes beyond checking for existence of aspect words in each sentence, by measuring the semantic similarity in the word embedding space. In order to calculate this factor for each sentence, we calculate the cosine similarity between a vector representing the sentence, and another vector representing the M aspect words. We use a pre-trained set of word vectors

from GloVe [23] and calculate the vector representations by averaging on the word vectors.

AEMI: The AEMI value of the most common bigram of a sentence indicates how often that aspect has been discussed among users. Thus, we use this value as a factor contributing to our SS2 score.

SS2 score is calculated based on weighted sum of all factors. We specify a weight for each factor to put more emphasize on some factors over others.

$$SS2 = w_{res}s_{res} + w_{cmp}s_{cmp} + w_{len}s_{len} + w_{snt}s_{snt} + w_{coh}s_{coh} + w_{cov}s_{cov} + w_{rat}s_{rat} + w_{rel}s_{rel} + w_{aemi}s_{aemi} \quad (2)$$

where w_i and s_i are the weight and factor value used for i^th factor, and res, cmp, len, snt, coh, cov, rat, rel, aemi represent reasoning, comparison, length, sentiment, coherence, coverage, rating and AEMI factors respectively. Finally, the sentences associated with each aspect are ordered in decreasing order of their SS2 score and sent to the categorization step.

3) Ranking and Categorizing Sentences: The sentences ordered by SS2 are then categorized into two groups of pros and cons based on their sentiment score. Specifically, first the sentiment score is calculated for each next sentence from the top of the ranked list. If the sentiment score falls in the sentiment threshold conditions the sentence is added to the corresponding list. In our experiments we used threshold of +0.4 and greater for pros, and -0.1 and smaller for cons. The process stops once both lists contain K sentences or when no sentences left. Finally, the pros and cons lists are returned.

Algorithm 3 findProsCons

Require: $aspectWords, aspectReviews, K$
  proSents = conSents = []
  for each $aspect_i$ **do**
    // Calculate AEMI to find the representative word pairs
    $aemi_i$ = AEMI($aspectWords_i$, $aspectReviews_i$)
    // select sentences that contain representative word pairs
    $sents_i$ = selRepSents($aemi_i$, $aspectReviews_i$)
    // Calculate significance of the sentences with SS2
    $rankedSents_i$ = rankBySS2($sents_i$, $aemi_i$)
    // Categorize the K most significant sentences into pros and cons
    $proSents_i$, $conSents_i$=categorize($rankedSents_i$, K)
    proSents.add($proSents_i$)
    conSents.add($conSents_i$)
  end for
  return proSents, conSents

C. Summarizing Aspect Words

In addition to the pros and cons identified by the last steps, we provide a summarized description of the aspects. Although the word pairs generated in Alg. 3 could be used to describe the aspects, the word pairs might not be maningful. For example, {"reservoir", "water"} is a pair but "water reservoir" would be more meaningful. We did not need to consider the word ordering in Alg. 3 because a descriptive/opinion word might be before or after the aspect word. Therefore, to find meaningful bigrams (not just pairs) we calculate AEMI of both bigrams in each pair (e.g. "reservoir water" and "water reservoir"). That is, we calculate an ordered AEMI table for the pairs (*word*, *aspectWord*) to take into account the ordering of the co-occurrence as well. As a result, for a bigram to score high, the two words should often co-occur in the same order of the bigram. The interpretation of each term in Eq. 1, in this case, is different from Sec. IV-B as the probability calculations take the ordering of the word occurrences into account. For example, $p(A, B)$ is the probability that A occurs before B in a sentence. Also, $p(A, \bar{B})$ is the probability that A exists but not followed by B, that is, if B occurs before A it is not counted. $p(A)$ is the same as before; the probability that a sentence contains A. We note that the AEMI table for this task, unlike the one in Sec. IV-B, is fully-filled (except for the diagonal).

After finding the representative bigrams for each aspect, we use a greedy approach to find the minimal set of bigrams that cover the selected pro and con sentences. Specifically, we order the bigrams in decreasing order of their AEMI values, select the first bigram, and eliminate all pro/con sentences that cover (contain) the bigram. The next bigram is selected only if there is a pro/con sentence that is not covered by previously selected bigrams. The process stops once all sentences are eliminated. Finally, the set of selected bigrams is returned. The overall process is illustrated in Alg. 4.

Algorithm 4 summarizeAspects

Require: $aspectWords, aspectReviews, proSents, conSents$
  aspectPhrases = []
  for each $aspect_i$ **do**
    //Calculate AEMI to find the representative bigrams
    $oaemi_i$ = OrderedAEMI($aspectWords_i$, $aspectReviews_i$)
    //Select the most representative bigrams
    $bigrams_i$ = selTopBigrams($oaemi_i$)
    //find the minimal set of bigrams to cover pro/conSents
    $minBigrams_i$ = findMinSet($bigrams_i$, $proSents$, $conSents$)
    aspectPhrases.add($minBigrams_i$)
  end for
  return aspectPhrases

V. Experimental Evaluation

A. Evaluation Criteria

We evaluate the effectiveness of pros and cons extracted by our technique compared to those of Kim and Hovy [19] that we call KH06 hereafter. Our evaluation criterion is the extent to which sentences selected by each algorithm indicate meaningful pros and cons. We calculate Discounted Cumulative Gain (DCG) to quantify the ordering quality of the sentences in the pros-and-cons table. We expect more relevant sentences to appear higher in the table. DCG for a pros-and-cons table is calculated as an average between DCG of pros and cons lists. $DCG(sentenceList) = \sum(\frac{rel_i}{log(i+1)})$ where $sentenceList$ can be $prosList$ or $consList$, and rel_i is relevance of sentence i.

The evaluation is performed at three levels: 1) Product Level: The pros and cons are considered relevant as long as they are about the product. 2) Aspect Level: The pros and cons should be related to the product and the given aspect. 3) Aspect Summarization: We measure the quality of the representative bigrams extracted by our aspect summarization algorithm. We use DCG to evaluate the quality of the pros and cons extracted by the algorithms in product level and aspect level. Plus, we use precision to evaluate the aspect summarization algorithm. Precision is calculated as the number of correctly identified bigrams divided by the total number of identified bigrams.

To establish the ground truth of the relevance of a sentence, we asked three evaluators who are graduate students in computer science and engineering fields, and are not authors of this paper, to label each of the outputted sentences by one of three classes pro, con, neither based on their personal opinion. The relevance of a sentence is one when the predicted class label agrees with the majority of the three opinions. If a majority vote cannot be established the sentence is considered as "neither". To avoid bias toward any algorithm, messages selected by the different algorithms were merged into one set before evaluation.

The evaluators were asked five questions to provide ground truth for three levels of evaluations. The first two questions were about the quality of the pros and cons, and the next three questions were related to the quality of aspect summarization task.

Given a list of top-K (K=10 in this case) pros and cons sentences, the first question aims to determine whether each sentence represented a pro, con, or neither. This question is at the product level and does not ask the evaluators to consider any product aspects. In the second question, given a list of top-K (K=5 in this case) sentences per aspect, the goal is to determine whether each sentence is a pro, con, or neither with respect to the given aspect. Therefore, this question is at the aspect level and the evaluators should consider whether each sentence is relevant to the aspect. For example, "the cellphone takes great photos." can be marked as a "pro" when the given aspect is "camera". But if the given aspect is "size" then the sentence should be marked as "neither".

Question three provides a list of aspect words, generated by LDA, for each aspect. Question four provides a list of representative bigrams extracted by Alg. 4 for each aspect. In questions three and four we ask "*Does each of the following lists represent/describe at least one product aspect? yes/no*". We establish the ground truth based on the majority votes from our three evaluators. We calculate precision based on the number of lists that contain product aspects according to the evaluators.

In question five, we provide them with a list of aspect words and the corresponding list of bigrams for comparison. We ask "*Which list is more effective/meaningful to represent one (or more) product aspects? The list on the left, the list on the right, or neither*". After collecting the answers from the evaluators, We break the ties by marking them "neither" to represent a mistake. For example, one evaluator votes for the list on the left, the second one votes for the list on the right, and the third one votes for neither. We consider this case as "neither". Next, we calculate the percentage of the time the evaluators preferred the representative bigrams over the aspect words.

B. Data

We used three data sets from Amazon Reviews data collected and published by McAuley [24], [25]:

- Cellphone data set: contains 837 reviews about a 5.0 inch Android smart phone with product ID (asin) "B0090AAOUW".
- Phone Case data set: contains 766 reviews about a Otterbox Defender Series Case for iPhone 4 & 4S with product ID (asin) "B005SUHPO6".
- Coffee Maker data set: contains 985 reviews about a Keurig k-cup coffee maker with product ID (asin) "B000AQPMHA".

Among the attributes available for each review document, we used *reviewText* to obtain the main body of the review and *overall* to obtain the overall rating of the product assigned by the user between 1 and 5. We removed the non-English tokens, URLs and emojis from the text.

We used Subjective Clues [15] to calculate the opinion-bearing word features for KH06 algorithm. Also we used the pros and cons data set created and used in [3] to train the KH06 algorithm.

C. Procedures

Since KH06 doesn't propose a solution to extract aspects, we perform the comparison between our algorithm and KH06 in two modes; product-level and aspect-level discussed as follows.

Table I
DCG RESULTS AT PRODUCT-LEVEL

Cellphone			
	Pros	Cons	Avg
ProCon-PL	15.09	13.13	14.11
KH06	14.05	10.57	12.31
Phone Case			
ProCon-PL	14.05	10.34	12.19
KH06	10.98	5.71	8.35
Coffee Maker			
ProCon-PL	15.09	15.09	15.09
KH06	15.09	1.11	8.10

Table II
DCG RESULTS AT ASPECT-LEVEL

Cellphone			
	Pros	Cons	Avg
ProCon	9.03	9.38	9.21
KH06-AL	2.75	3.13	2.94
Phone Case			
ProCon	9.09	8.55	8.82
KH06-AL	2.68	2.67	2.67
Coffee Maker			
ProCon	8.77	8.89	8.83
KH06-AL	2.19	1.63	1.91

1) Product-Level: We simplify our algorithm to extract product-level pros and cons. Therefore, the aspect extraction step is removed from our algorithm. We refer to this version of our algorithm as "ProCon-PL". The ground truth label for each sentence in this mode is obtained from the first question from the evaluators as discussed in V-A. We used the labeled data set created by Liu et al. [3] to train the KH06 algorithm. The labels in the data set are either "pro" or "con". Next, we used the Amazon reviews data as the test data set. We collected top 10 pros and cons predicted by the algorithm in decreasing order of the prediction values (between 0 and 1) generated by the model.

2) Aspect-Level: In order to add aspects to KH06, we first extract aspects by our aspect extraction Alg. 2. Then for each aspect and its associated sentences, we run KH06 to find top-K pros and cons (K=5 in this case). We refer to this version of KH06 as "KH06-AL". The ground truth for each sentence in this mode is obtained from the second question from the evaluators as discussed in V-A.

3) Aspect Summarization: Finally, we evaluate the quality of the summarized aspects generated by Alg. 4. The ground truth is based on the fifth question from the evaluators as discussed in V-A.

D. Results on Pros and Cons

At the product-level, Table I illustrates the DCG results generated by KH06 compared to the product level version of our algorithm (ProCon-PL) discussed in V-C1 on the three data sets explained in V-B. Overall, ProCon-PL outperforms KH06 in product mode on all three data sets. The "Avg" column shows the average between the DCG results of pros and cons.

At the aspect-level, Table II shows the DCG results generated by KH06 with aspect extraction (KH06-AL) discussed in V-C2 and our algorithm (ProCon) on the three data sets explained in V-B. ProCon significantly outperforms KH06-AL in on all three data sets. The "Avg" column shows the average between the DCG results of pros and cons.

Examples of Pros and Cons Tables: Tables III and IV show example pros and cons from one example aspect identified by our algorithm (ProCon) compared to those of KH06 in aspect level (KH06-AL) on Coffee Maker data set. The top-5 aspect words that describe this aspect are *water, unit, tank, reservoir, noise*. By looking at the aspect words, one can realize that the aspect is about the water reservoir, tank and other related functionalities as well as effects such as noise. Therefore, it's desired to find pros and cons that are related to this aspect.

KH06-AL shows many mistakes mostly because the selected sentences are not related to the aspect. For example the first and second detected pros in Table III are pros, but they are not related to the aspect. Although the algorithm is using the aspect extraction Alg. 2 from our ProCon, it is still making many mistakes. The main reason is that the reviews assigned to aspects can be discussing multiple aspects or no aspects in different sentences. As a result, the sentences under an aspect may not be closely related to the aspect. This is where our SS2 method comes handy because, using its factors, it ensures the selected sentences to be closely related to the aspect in addition to be highly likely to represent pros or cons.

In addition to pros and cons, our ProCon provides a summary of each identified aspects in form of bigrams. The first row in Table IV shows the summary. First, the summary bigrams tend to point to aspects of the product that are very related together. They also indicate the key product aspects that are mentioned in the selected reviews. Therefore, by looking at them, one can quickly learn about the aspect and the ideas described by the selected sentences. Furthermore, in contrast to the unigram aspect words generated by LDA, the bigrams tend to be more meaningful and informative. For example, "unit" in the aspect words indicates a very broad concept, but when it is paired with "noise" the reader gets a better idea about a potential negative effect of the coffee maker.

Although our ProCon can find related pros and cons, there are instances that it makes mistakes. For example, the 4th and 5th pro and the 5th con in Table IV. The positive point is that the mistakes tend to occur lower in the list. Similar to the results from other aspects or other data sets, the mistakes tend to be "N" (neither) which mostly occur in complex situations where even though the aspect is discussed, what shapes the positive or negative opinion is not about the aspect. For example, in sentence 5 in cons column, the main reason for coffee losing its taste is the "thin plastic coffee cup" not the "hot water". One possible solution would be to

Table III
KIM2006-AL: EXAMPLE PROS & CONS FOR ASPECT WORDS (WATER,UNIT,TANK,RESERVOIR,NOISE) FROM COFFEE MAKER DATA SET

	Pros		Cons	
	Representative Sentence	GT	Representative Sentence	GT
1	easy to use and makes great coffee.	N	emailed customer support how to drain the unit for long term storage, was told there is no way.	C
2	great price, easy to use, wonderful espresso.	N	[...] there is no way to fully drain the coffeemaker of water.	C
3	the water tank is removable for cleaning and has clearly marked water level indicators.	P	i looked online for ways to drain the maker to get it off of my counter while using my espresso machine.	N
4	the convenience is great, and the coffee is good .	N	i went to keurig's website and learned that you can not drain the internal water reservoir.	C
5	this coffee maker makes quick, excellent single cups of coffee , which is great for small households.	N	if you forget to turn the power button on , before adding water and your k-cup , the water will drain into the machine.	C

Table IV
PROCON: EXAMPLE PROS & CONS FOR ASPECT WORDS (WATER,UNIT,TANK,RESERVOIR,NOISE) FROM COFFEE MAKER DATA SET

Summary	water reservoir, hot water, water heater, unit noise			
	Pros		Cons	
	Representative Sentence	GT	Representative Sentence	GT
1	nice thing [...] it was fast, as it pushed the hot water through with a good amount of pressure.	P	i have had no real problems with my b40 coffee maker until i cleaned the water reservoir.	C
2	the water reservoir is large enough for may daily coffee consumption.	P	be sure that you do not block the water outlets in the needle or you will get less coffee in the cup.	C
3	i had emptied the removable water reservoir and easily cleaned the appropriate parts.	P	you have to add more water to the water tank that already has enough to make a cup of coffee.	C
4	again, an add-in paper filter would be useful to slow the water and create a better brew slurry.	N	It worked fine for two weeks. Then it started making a horrible noise.	C
5	the energy efficiency paradigm with this coffee maker will likely be similar to that of on-demand water heaters	N	the coffee also loses its taste because hot water is pouring through thin plastic coffee cups	N

improve the Reasoning Factor of SS2 such that it identifies the cause and effect. In this example, the first sentence "The coffee also loses its taste" would be the effect, and "hot water is pouring through thin plastic coffee cups" would be the cause. Identifying causality in general is a difficult task, but perhaps it can be applied to sentences with explicit reasoning like the one discussed.

The next mistake is the fourth pro which is voted as "neither" by the evaluators. Unlike the previous mistake, where the sentence represents an actual con, but it is not due to a flaw of the aspect, this mistake does not seem to be a pro. The main reason that this sentence shows so high in the list is the high SS2 score which is due to positive words like "useful" and "better" that lead to high sentiment value. One possible way to remedy this case is to take into account the grammar relationship (e.g. by dependency tree analysis) and assign higher importance to words in closer relationship with the aspect words. The last mistake is in the fifth pro. This is also different from the last two mistakes because the sentence can be a pro or a con depending on whether or not the author considers "on-demand water heaters" to be efficient. The sentence has been surfaced in the list of pros because of multiple factors of SS2 like comparison, coverage of the aspect words, high coherence, high relevance, high sentiment, and high rating of the review the sentence belongs to. Sentiment has the highest weight in SS2 equation and it is 0.56 in scale of -1 to 1 for this sentence. The main words that contribute to the sentiment score are "efficiency" and "likely". It seems that the sentiment for this sentence is overestimated. One possible solution is to improve the sentiment analysis algorithm.

Table V
PRECISION OF ASPECTS: THE FIRST COLUMN SHOWS THE PRECISION OF ASPECT WORDS, AND THE SECOND COLUMN SHOWS THE PRECISION OF REPRESENTATIVE BIGRAMS, BETWEEN ZERO AND ONE.

	Aspect Word	Representative Bigram
Cellphone	0.43	1.0
Phone Case	0.25	0.75
Coffee Maker	0.60	1.0

E. Results on Aspect Summarization

Table V shows that the representative bigrams have a higher precision than aspect words in identifying aspects based on the three data sets. The precision values indicate the ratio of aspect words or bigrams that represent informative aspects of the product based on the user evaluation. The precision results show that unlike the aspect words, the bigrams extracted by Alg. 4 are likely to represent a meaningful aspect of the product.

Table VI shows a direct comparison between aspects represented by representative bigrams and single words (aspect words) for our three data sets. The values show the percentage distribution of the evaluators' preference on the bigram aspects, single word aspects, and neither. We can see that at least 75% of the time the aspects represented by bigrams are preferred as more informative.

Table VI
ASPECT WORDS VS. BIGRAMS: SHOWS THE PERCENTAGE THE EVALUATORS' PREFERENCE DISTRIBUTION ON THE REPRESENTATIVE BIGRAMS, ASPECT WORDS AND NEITHER.

	% Preferred		
	Representative Bigrams	Aspect Words	Neither
Cellphone	86	0	14
Phone Case	75	25	0
Coffee Maker	80	20	0

Table VII
ASPECT SUMMARIZATION ON COFFEE MAKER DATA SET - THE C (I.E. CONTAINS) COLUMNS SHOW THE WHETHER EACH LIST CONTAINS AT LEAST ONE PRODUCT FEATURE BASED ON THE EVALUATORS' OPINION, AND THE P (I.E. PREFERRED) COLUMN SHOWS THE PREFERRED LIST.

List-1: Aspect Words	C	List-2: Bigrams	C	P
size, elite, brew, model, mug	✓	cup size, expensive elite, Krups model, easy brew	✓	List-1
quality, press, french, milk, tons		coffee quality, french coffee	✓	List-2
coffee, cup, machine, make, love	✓	coffee cup, hot coffee	✓	List-2
water, unit, tank, reservoir, noise	✓	water reservoir, hot water, water heater, unit noise	✓	List-2
months, machine, problem, service, customer		customer service, customer support, common problem	✓	List-2

Examples of Aspect Summarization: Table VII shows results on aspect summarization task on Coffee Maker data set. The second column depicts the top aspect words extracted by LDA, and the fourth column shows the aspect summaries (i.e. the minimum set of representative bigrams) extracted via Alg. 4. Each method of summarization has the information about the evaluators' preference in the GT column on its right. Overall, the table shows that our algorithm finds five different aspects for the coffee maker product and provides a summary for each aspect based on the reviews left by the users. Moreover, the aspects that are expressed in bigram form tend to be more meaningful and informative compared to the single words generated by LDA in Alg. 2. The only instance where the aspect words are preferred over the bigrams is on the first aspect. It has been voted two to one for the aspect word list by the evaluators. "cup size" could be considered as a product aspect, indicating the number of cups each brew can produce, or it can be the coffee cup which is not related to the coffee maker machine. Also, "easy brew" is a feature of the specific coffee maker model as explained in the product specifications. However, it may not have seemed like a clear product aspect for the evaluators. On the aspect word list "size" seems to be the only word that could be considered as an aspect for the product, which is ambiguous because it can be applied to different parts of the coffee maker.

Table. VIII shows results on aspect summarization task on Cellphone data set. Similar to Table. VII, the second column depicts the top aspect words extracted by LDA, and the fourth column shows the aspect summaries (i.e. the minimum set of representative bigrams) extracted via Alg. 4. Each method of summarization has the information about the evaluators' preference in the GT column on its right. Overall, the table shows that our algorithm finds seven different aspects for the coffee maker product and provides a summary for each aspect based on the reviews left by the users. Moreover, the aspects that are expressed in digram form tend to be more meaningful and informative compared to the single words generated by Alg. 2. The only instance where bigrams were not preferred is on the second aspect. Although the bigrams contain at least a product aspect, according to the evaluators, neither of the representations are preferred. The AEMI value for "signal bars" is about 0.05 which is not significant but the total number of reviews for this aspect is small, and this bigram was selected to represent the aspect. One possible way to remedy this situation is to exclude aspects with small number of reviews from the results of Alg. 2 because statistical methods like AEMI tend to be more reliable with larger number of instances.

Table VIII
ASPECT SUMMARIZATION ON CELLPHONE DATA SET - THE C (I.E. CONTAINS)COLUMNS SHOW THE WHETHER EACH LIST CONTAINS AT LEAST ONE PRODUCT FEATURE BASED ON THE EVALUATORS' OPINION, AND THE P (I.E. PREFERRED) COLUMN SHOWS THE PREFERRED LIST.

List-1: Aspect Words	C	List-2: Bigrams	C	P
back, money, buy, screen	✓	money back, money saving, battery back	✓	List-2
card, talk, data, service, plan		signal bars	✓	Neither
battery, time, box, screen	✓	screen protector, battery charger	✓	List-2
good, screen, love, price		screen crack, large screen	✓	List-2
manual, year, instruction, text		manual size, text message	✓	List-2
memory, storage, low, card, space	✓	internal memory, internal space	✓	List-2
camera, jack, note, work		front camera, play music, camera flash	✓	List-2

VI. CONCLUSION

We extract aspects from product reviews and assign the review sentences to the related aspects. We introduce a modified version of Significance Score (SS2) with additional factors to find sentences that are both likely to represent pros and cons, and closely related to the aspect they belong to. Finally, we provide a a summary for each product aspect in the form of bigrams such that each bigram shows a description or opinion about the aspect.

REFERENCES

[1] M. Hu and B. Liu, "Mining opinion features in customer reviews," in *AAAI*, vol. 4, no. 4, 2004, pp. 755–760.

[2] Z. Hai, K. Chang, and J.-j. Kim, "Implicit feature identification via co-occurrence association rule mining," in *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, 2011, pp. 393–404.

[3] 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*. ACM, 2005, pp. 342–351.

[4] Y. Zhao, B. Qin, S. Hu, and T. Liu, "Generalizing syntactic structures for product attribute candidate extraction," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2010, pp. 377–380.

[5] N. Jakob and I. Gurevych, "Extracting opinion targets in a single-and cross-domain setting with conditional random fields," in *Proceedings of the 2010 conference on empirical methods in natural language processing*. Association for Computational Linguistics, 2010, pp. 1035–1045.

[6] H. Lakkaraju, C. Bhattacharyya, I. Bhattacharya, and S. Merugu, "Exploiting coherence for the simultaneous discovery of latent facets and associated sentiments," in *Proceedings of the 2011 SIAM international conference on data mining*. SIAM, 2011, pp. 498–509.

[7] T. Nasukawa and J. Yi, "Sentiment analysis: Capturing favorability using natural language processing," in *Proceedings of the 2nd international conference on Knowledge capture*. ACM, 2003, pp. 70–77.

[8] N. Kobayashi, R. Iida, K. Inui, and Y. Matsumoto, "Opinion mining on the web by extracting subject-aspect-evaluation relations." in *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, 2006, pp. 86–91.

[9] C. H. E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *Eighth International Conference on Weblogs and Social Media (ICWSM-14). Available at (20/04/16) http://comp. social. gatech. edu/papers/icwsm14. vader. hutto. pdf*, 2014.

[10] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proceedings of the 2013 conference on empirical methods in natural language processing*, 2013, pp. 1631–1642.

[11] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*. Association for Computational Linguistics, 2002, pp. 79–86.

[12] P. D. Turney, "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews," in *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2002, pp. 417–424.

[13] J. M. Wiebe, R. F. Bruce, and T. P. O'Hara, "Development and use of a gold-standard data set for subjectivity classifications," in *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*. Association for Computational Linguistics, 1999, pp. 246–253.

[14] E. Riloff, J. Wiebe, and T. Wilson, "Learning subjective nouns using extraction pattern bootstrapping," in *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*. Association for Computational Linguistics, 2003, pp. 25–32.

[15] E. Riloff and J. Wiebe, "Learning extraction patterns for subjective expressions," in *Proceedings of the 2003 conference on Empirical methods in natural language processing*. Association for Computational Linguistics, 2003, pp. 105–112.

[16] Y. Lu, C. Zhai, and N. Sundaresan, "Rated aspect summarization of short comments," in *Proceedings of the 18th international conference on World wide web*. ACM, 2009, pp. 131–140.

[17] Z. Hai, G. Cong, K. Chang, W. Liu, and P. Cheng, "Coarse-to-fine review selection via supervised joint aspect and sentiment model," in *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. ACM, 2014, pp. 617–626.

[18] E. Ahmadzadeh and P. K. Chan, "Mining pros and cons of actions from social media for decision support," in *Big Data (Big Data), 2017 IEEE International Conference on*. IEEE, 2017, pp. 877–882.

[19] S.-M. Kim and E. Hovy, "Automatic identification of pro and con reasons in online reviews," in *Proceedings of the COLING/ACL on Main conference poster sessions*. Association for Computational Linguistics, 2006, pp. 483–490.

[20] H.-r. Kim and P. K. Chan, "Identifying variable-length meaningful phrases with correlation functions," in *Tools with Artificial Intelligence, 2004. ICTAI 2004. 16th IEEE International Conference on*. IEEE, 2004, pp. 30–38.

[21] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[22] N. Jindal and B. Liu, "Mining comparative sentences and relations," in *AAAI*, vol. 22, 2006, pp. 1331–1336.

[23] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543. [Online]. Available: http://www.aclweb.org/anthology/D14-1162

[24] M. Wan and J. McAuley, "Modeling ambiguity, subjectivity, and diverging viewpoints in opinion question answering systems," in *Data Mining (ICDM), 2016 IEEE 16th International Conference on*. IEEE, 2016, pp. 489–498.

[25] J. McAuley and A. Yang, "Addressing complex and subjective product-related queries with customer reviews," in *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2016, pp. 625–635.