Identifying Usage Expression Sentences in Consumer Product Reviews

Shibamouli Lahiri and V. G. Vinod Vydiswaran and Rada Mihalcea
University of Michigan
{lahiri, vgvinodv, mihalcea}@umich.edu

Abstract

In this paper we introduce the problem of identifying usage expression sentences in a consumer product review. We create a human-annotated gold standard dataset of 565 reviews spanning five distinct product categories. Our dataset consists of more than 3,000 annotated sentences. We further introduce a classification system to label sentences according to whether or not they describe some “usage.” The system combines lexical, syntactic, and semantic features in a product-agnostic fashion to yield good classification performance. We show the effectiveness of our approach using importance ranking of features, error analysis, and cross-product classification experiments.

1 Introduction

Identification of usage expressions — phrases or sentence snippets describing product use in reviews — is an important problem in mining consumer product reviews. Identifying such usage expressions accurately allows us to view the relationship between consumers and products more clearly (e.g., by indicating how frequently a consumer uses a product). Further, the language and style employed in describing product use bring relevant and unseen aspects of the products to the fore (e.g., describing usage of a product in non-traditional and unique ways).

Usage expressions can take several forms, such as which aspects of the product are used, why the product is used, how it is used, when it is used, and so forth (c.f. Section 3 for specific examples). The product could be used by a consumer in a number of ways, sometimes in unique ways not intended for originally. Hence enumerating all possible uses of a product is computationally intractable. In this paper, therefore, we focus on four specific cases of product usage: why the product is used, where it is used, how it is used, and if there are any non-standard or non-traditional use (cf. Section 3).

While the relationship between product usage and consumer behavior has mostly been discussed by marketing researchers and psychologists, the question of whether the phenomenon of usage has any detectable signature in terms of the language used by consumers has not been addressed thus far. In this paper, we introduce the task of identifying usage expressions from consumer product reviews. In particular, we focus on classifying review sentences as to whether they contain a usage expression or not. We create our own human-annotated corpus of 565 reviews on five distinct product categories containing more than 3000 sentences. We introduce a system that classifies sentences according to whether they contain a usage expression or not. We create our own human-annotated corpus of 565 reviews on five distinct product categories containing more than 3000 sentences. We introduce a system that classifies sentences according to whether they contain a usage expression or not with 87.2% accuracy. We also show that an appropriate combination of lexical, syntactic, and semantic features performs better than individual feature categories.

2 Related Work

Existing research could be organized into six self-consistent psycho-sociological theories, namely psycho-analysis, social theories, stimulus-response theories, trait and factor theories, self-theories, and life style theories. Kassarjian (1971) offers a comprehensive review of the literature on consumer behavior and psychological traits. Robertson and Myers (1969) found weak relationships between opinion leadership and innovative buying behavior, but observed that the relationship strength varied by product category. Tucker and Painter (1961), and Sparks and Tucker
(1971) showed that there were correlations between personality traits and the types of products used. Dolich (1969) posited that products as symbols were organized into congruent relationships with the consumer’s self-image. More recently, Govers and Schoormans (2005) found that people preferred products with a product personality that matched their self-image, and the positive effect of product-personality congruence was independent of user-image congruence.

In natural language processing research, the closest problem to usage expressions is perhaps that of opinion mining from product reviews and product aspects. Dave et al. (2003) classified reviews as expressing positive or negative sentiment. They identified four problems with review classification, including rating inconsistency, ambivalence, data sparseness, and skewed distribution. Hu and Liu (2004) extracted product features from the reviews of a single product, taking user opinion into account. Opinion/product features were mined if a reviewer had commented on them. Popescu and Etzioni (2005) presented OPINE, an unsupervised information extraction system that mined reviews in order to build a model of important product features, their evaluation by reviewers, and their relative quality across products. OPINE’s use of relaxation labeling led to strong performance on the tasks of finding opinion phrases and their polarity. Ding et al. (2008) presented a “holistic lexicon-based approach” for mining context-dependent opinion words. The proposed method used an aggregating function for multiple conflicting opinion words in a sentence. The authors further implemented a system called “Opinion Observer” based on their method. Lastly, Wu et al. (2009) implemented a special dependency parser for opinion mining that used phrases (rather than words) as the primitive building blocks. Since many product features are in fact phrases, this approach led to good results for extracting relations between product features and opinion expressions.

Yet another related task is that of mining semantic affordances (Chao et al., 2015). In this task, “usage” of a product can be viewed as an action performed on an object with the help of the product. Relationships between such actions and objects are known as “semantic affordances”. As Chao et al. showed, text mining can be very effective at ascertaining affordance relationships between verb and noun classes. Similar verb-noun relationships have also been formulated in the problem of learning selectional preferences from text (Resnik, 1997; Brockmann and Lapata, 2003; Erk, 2007; Pantel et al., 2007; Bergsma et al., 2008; Van de Cruys, 2014), and more generally, in the problem of probabilistic frame induction (Chambers and Jurafsky, 2011; Cheung et al., 2013; Chen et al., 2013).

Another topic of research related to our work is the problem of research idea extraction from academic papers. Gupta and Manning (2011) took the first stab at this problem by implementing a bootstrapping algorithm on dependency tree kernels. Gupta and Manning’s method was later refined by Tsai et al. (2013) who worked with a more crisp set of idea categories. We view this problem as conceptually parallel to ours; however, a key difference is that usage expressions are typically more obscure in text as compared to research ideas.

3 Building a Usage Expression Dataset

Product reviews often contain usage information. Specifically, in addition to opinions on product quality, reviewers often share how, where, or why they use the product. We therefore build our dataset of product usage expressions starting with a collection of product reviews.

We collect Amazon product reviews for five different product categories, as shown in Table 1. The particular product lines we use are: a laundry product: specifically, Downy Unstopables In Wash Fresh Scent Booster 13.2 Oz; two kinds of cooking agents, namely, Olive oil: Baja Precious Extra Virgin Olive Oil from Baja California (750ml Bottle) and Vinegar: Raw Organic Apple Cider Vinegar by Bragg (1 gallon); a Medicine: Kirkland Signature Low Dose Aspirin, 2 bottles – 365-Count Enteric Coated Tablets each; and a household item, namely Toothpaste: Colgate Optic White Toothpaste, 4 Ounce (Pack of 2). The reviews are split into sentences, with the total number of sentences and average number of sentences per review as shown in Table 1. In all, there are 3020 sentences in 565 reviews, with an average of 5.34 sentences per review.

With the help of three linguistics undergraduate students, each sentence in the dataset was annotated as containing a usage expression or not. Initially, as an early trial, we asked the annota-
### Table 1: Product categories in our dataset.

<table>
<thead>
<tr>
<th>Product category</th>
<th>Product</th>
<th># Reviews</th>
<th># Sentences</th>
<th>Avg # Sentences per Review</th>
</tr>
</thead>
<tbody>
<tr>
<td>Laundry product</td>
<td>Scent booster</td>
<td>125</td>
<td>695</td>
<td>5.56</td>
</tr>
<tr>
<td>Cooking agent</td>
<td>Olive oil</td>
<td>110</td>
<td>588</td>
<td>5.35</td>
</tr>
<tr>
<td>Cooking agent</td>
<td>Vinegar</td>
<td>110</td>
<td>623</td>
<td>5.66</td>
</tr>
<tr>
<td>Medicine</td>
<td>Aspirin</td>
<td>110</td>
<td>463</td>
<td>4.21</td>
</tr>
<tr>
<td>Household item</td>
<td>Toothpaste</td>
<td>110</td>
<td>651</td>
<td>5.92</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td><strong>565</strong></td>
<td><strong>3020</strong></td>
<td><strong>5.34</strong></td>
</tr>
</tbody>
</table>

In the first step, we instructed the annotators to read each product review carefully, identify all usage expressions in the review (examples below), and write them in a given textbox, one usage expression per line. Annotators were requested to write the usage expressions in their own words. This component was employed to make sure annotators carefully read and understood the review.

The second step involved answering the following four questions on usage types:

1. **(A)** Does the sentence describe why the product was being used? (usage reason/purpose) E.g., “I used unstopables to freshen my room.”

2. **(B)** Does the sentence describe where the product was used? E.g., “I used unstopables with my cat litter.”

3. **(C)** Does the sentence describe how the product was used? E.g., “I use three cups of Downy Unstopables in every wash.”

4. **(D)** Does the sentence describe any non-traditional or non-standard usage of the product? E.g., “I always love to add some hot water to unstopables and make my own DIY air freshener!”

If a sentence had a positive answer to one or more of these four questions, then it was labeled as containing a usage expression.

Additionally, several specific instructions were added to deal with potentially difficult or complex cases, by asking annotators to (1) consider the context (one sentence before and after the target sentence) before deciding whether to mark a sentence or not. (2) determine if a sentence contains an opinion (“Love it”, “Hate it”, etc.) or a recommendation (“I’d recommend this product to all aspiring gardeners”), and if so, pairing it with an explicit usage expression in some form. (3) determine if a sentence talks about usage of another product that is not the primary focus of the review (i.e., a secondary product), then mark the sentence only if the primary product is being used in addition to the secondary product. (4) determine if the secondary product is used instead of the primary product: “Unstopables were not good, so I used sheets instead.”, or if only the secondary product was used: “I used sheets, they are better.” then do not label the sentence. (5) focus only on products, and ignore other (named) entities like persons, organizations, locations, and dates.

Table 2 shows an example product review, and sentences that were agreed upon by all annotators to contain, or not, a usage expression. We also show sentences on which there was no consensus. Note that such sentences have a fair amount of ambiguity. For example, the sentence “I do recommend this for times when you may want extra freshness for your clothes or towels.” does not seem to contain an explicit usage expression, but it does indicate that the consumer used the product to obtain extra freshness for clothes or towels. Sentences like this demonstrate the difficulty of identifying usage expressions in product reviews.

Inter-annotator agreement values, shown in Table 3, indicate that the task is moderately difficult. We can see that different products have different difficulty levels, with Vinegar being the least difficult (highest $A_3$ agreement as well as highest $\kappa$), while for the other four products, $\kappa$ was between 0.43 and 0.48. This is presumably owing to the fact that Vinegar is a cooking agent and used in many different ways, thus providing more opportunity to find a usage sentence (by several people) in a product review.

To construct a gold standard, we took the majority of the three votes assigned by the three anno-
Sample Review
I used this recently when I washed my blankets and towels, and I was definitely impressed. Just a small amount (half a capful) was necessary to give my blankets and towels an extra burst of freshness. The scent is a little bit floral and lasts for a few days. I put the Downy booster directly into the washer. (Instructions say NOT to put in your dispenser) And it does work fine with high efficiency washers. I do recommend this for times when you may want extra freshness for your clothes or towels.

Usage annotations (agreed by all)
I used this recently when I washed my blankets and towels, and I was definitely impressed. Just a small amount (half a capful) was necessary to give my blankets and towels an extra burst of freshness.

Non-usage annotations (agreed by all)
The scent is a little bit floral and lasts for a few days.

Mixed usage/non-usage annotations
I put the Downy booster directly into the washer. (Instructions say NOT to put in your dispenser) And it does work fine with high efficiency washers. I do recommend this for times when you may want extra freshness for your clothes or towels.

Table 2: An example review and its annotations.

4 Finding Usage Expression Sentences
Once the annotated dataset was finalized, our primary goal was to build a classifier to predict if a given sentence contains usage expressions or not. We learn the classifier over five categories of features extracted from the sentence and neighboring context. In this paper, we show the performance on a small development dataset using a logistic regression classifier, chosen based on its performance on a small development dataset of usage-annotated sentences drawn from 20 product reviews. The following features are included:

(A) Lexical features: As n-grams are usually very helpful in document classification, we explore their utility on the task of usage expression sentence classification. We use word unigrams and bigrams, part-of-speech (POS) bigrams, and character trigrams. We use the CRFTagger (Phan, 2006) for POS tagging.

(B) Embeddings: Embeddings encode latent semantics and could reflect usage patterns. We train a word embedding using word2vec (Mikolov et al., 2013) over a large corpus of 55,463 product reviews. This corpus is constructed from all Amazon reviews associated with any product that has “Unstopables”, “Olive oil”, “Vinegar”, “Aspirin”, or “Toothpaste” in its title. Once the word embedding is trained, a sentence is represented by the weighted average of the embeddings of all the unique words in it.

(C) Syntax: We use bags of constituency and dependency production rules, obtained from the output of the Stanford parser (Klein and Manning, 2003; Chen and Manning, 2014). For constituency grammar, we use terminal and non-terminal rules separately as well as together. For the dependency grammar, we use the (collapsed) dependency types (amod, nsubj, etc.), and the lexicalized dependencies (e.g., (nsubj, Kirkland, seems)) as separate features.

(D) Style: We extract thirteen shallow surface-level and style features to encode the stylistic properties of a sentence, in the hope that they would be predictive of whether the sentence contains a usage expression. These features are: sentence position, average word length (in chars), sentence length (in words and characters), type-token ratio, Flesch Reading Ease (Flesch, 1948; Farr et al., 1951), Automated Readability Index (Senter and Smith, 1967), Flesch-Kincaid Grade Level (Kincaid et al., 1975), Coleman-Liau Index (Coleman and Liau, 1975), Gunning Fog Index (Gunning, 1968), SMOG Score (McLaughlin, 1969), Formality (Heylighen and Dewaele, 1999), and Lexical Density (Ure, 1971).

(E) Semantics: Since usage is above all a semantic phenomenon, a semantic space should be able to capture the dominant properties of the usage expression. We use the following feature sets to capture a semantic space for a sentence. Each feature set effectively describes a lexicon, and we turn “on” the features in the lexicon that are present in the target sentence.

1. Product categories: This feature set consists of the list of product categories obtained from the Walmart API.2 We use both main categories and sub-categories.

2. Concreteness: The set of words, along with their concreteness scores, available as part of the Free Association Norms Database (Nelson et al., 1998). There are more than 3,000 words available as part of the database.

3. Levin classes: The set of coarse and fine-grained variations of Levin verb classes and

2https://developer.walmartlabs.com/
verb alternations, leading to four types of features (Levin, 1993).

4. LIWC: Like Levin classes, we included another set of features derived from the LIWC dictionary of psychological word categories (Tausczik and Pennebaker, 2010).

5. Semantic lexicons: Like Levin classes, we use the Roget thesaurus and WordNet Affect word categories, with a binary feature representation. If a word falls under any of the Roget word categories, the corresponding feature is set.

6. Named Entities: We use the Stanford NER (Finkel et al., 2005) to identify named entities in our corpus, and then use these entities as bag-of-features. We use the terms, the entity types, and the lexicalized entity types (terms + entities) as our bags. Standard tf, tfidf, and binary representations are used. We use the seven-class typology of named entities (Location, Person, Organization, Money, Percent, Date, Time).

7. Spatial Prepositions: Recent studies have shown prepositions to be a precious source of semantic information (Srikumar and Roth, 2013; Schneider et al., 2015, 2016). We use a lexicon of spatial prepositions as a bag-of-words feature. The rationale was to observe if spatial properties of usage of objects (“use olive oil with celery”, “put detergent in washer”) can be captured in terms of prepositions such as on, in, by, with, etc.

8. Semantic Distance: Finally, we added the (weighted) WordNet distance between all words and the verb use, where weights are set as binary, tf, and tfidf, as before. The rationale behind this feature is that it captures words similar to the verb use in the sentence, and their relative importance.

5 Evaluation

We use the dataset introduced in Section 3 to evaluate the accuracy of the usage detection classifier. 20% of the data for each product is held out as test data, and the remaining 80% is used for training.

We start by evaluating each individual feature using a ten-fold cross-validation on the training data. We then explore three combination methods, applied on a subset of seven feature sets, selected based on their performance and diversity: word unigrams, POS bigrams, character trigrams, embeddings, constituency rules, product categories, and concreteness. We combine these features through: classifier voting, where we assign the class predicted by the majority of the classifiers; feature fusion, where we join all the individual features into one feature vector used in the classification; and meta-learning, where we use the output of the individual classifiers as input into another classifier (again using logistic regression for the meta-learner). Table 4 shows the results of these evaluations. As seen in the table, while simple features, such as word n-grams and character trigrams, lead to the best performance among the individual features, better performance is obtained when they are combined with other features (bottom rows of Table 4).

The meta-learner based combination strategy resulted in the best performing classifier during the cross-validation experiments on training data. We next evaluate this classifier on the test data consisting of 20% reviews of all five products. Table 5 shows the results obtained on the test data. For

---

Table 3: Majority label statistics, and three-way inter-annotator agreement. $A_3$ is the % of sentences where all three annotators agreed. $\kappa$ is the Fleiss’ kappa among three annotators (Fleiss, 1971).

<table>
<thead>
<tr>
<th>Product type</th>
<th>Majority Yes</th>
<th>Majority No</th>
<th>Majority Not Sure</th>
<th>All Yes</th>
<th>All No</th>
<th>$A_3$</th>
<th>$\kappa$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scent booster</td>
<td>201</td>
<td>494</td>
<td>0</td>
<td>80</td>
<td>385</td>
<td>66.91</td>
<td>0.46</td>
</tr>
<tr>
<td>Olive oil</td>
<td>91</td>
<td>493</td>
<td>4</td>
<td>40</td>
<td>395</td>
<td>73.98</td>
<td>0.43</td>
</tr>
<tr>
<td>Vinegar</td>
<td>190</td>
<td>430</td>
<td>3</td>
<td>139</td>
<td>369</td>
<td>81.54</td>
<td>0.71</td>
</tr>
<tr>
<td>Aspirin</td>
<td>94</td>
<td>366</td>
<td>3</td>
<td>47</td>
<td>282</td>
<td>71.06</td>
<td>0.48</td>
</tr>
<tr>
<td>Toothpaste</td>
<td>137</td>
<td>514</td>
<td>0</td>
<td>56</td>
<td>411</td>
<td>71.74</td>
<td>0.46</td>
</tr>
<tr>
<td>Overall</td>
<td>713</td>
<td>2297</td>
<td>10</td>
<td>362</td>
<td>1842</td>
<td>72.98</td>
<td>0.52</td>
</tr>
</tbody>
</table>

---


5. We use the Wu-Palmer similarity (Wu and Palmer, 1994).
### Table 4: Micro-averaged sentence-level results (%) under 10-fold cross-validation on the training data. Maximum value in each column (within each section) is boldfaced.

<table>
<thead>
<tr>
<th>Feature Type</th>
<th>Prec.</th>
<th>Rec.</th>
<th>F-score</th>
<th>Accu.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word unigrams</td>
<td>71.56</td>
<td>54.94</td>
<td>62.16</td>
<td>83.88</td>
</tr>
<tr>
<td>Character trigrams</td>
<td>70.06</td>
<td>57.19</td>
<td>62.98</td>
<td>83.80</td>
</tr>
<tr>
<td>POS bigrams</td>
<td>55.72</td>
<td>39.69</td>
<td>46.36</td>
<td>77.87</td>
</tr>
<tr>
<td>Embeddings</td>
<td>71.92</td>
<td>47.49</td>
<td>57.20</td>
<td>82.88</td>
</tr>
<tr>
<td>Constituency</td>
<td>70.49</td>
<td>52.17</td>
<td>59.96</td>
<td>83.22</td>
</tr>
<tr>
<td>Dependency</td>
<td>57.53</td>
<td>33.10</td>
<td>42.02</td>
<td>78.00</td>
</tr>
<tr>
<td>Style</td>
<td>54.17</td>
<td>11.27</td>
<td>18.65</td>
<td>76.33</td>
</tr>
<tr>
<td>Product categories</td>
<td>67.19</td>
<td>44.37</td>
<td>53.44</td>
<td>81.38</td>
</tr>
<tr>
<td>Concreteness</td>
<td>59.61</td>
<td>53.21</td>
<td>56.23</td>
<td>80.04</td>
</tr>
<tr>
<td>Levin classes</td>
<td>59.72</td>
<td>37.26</td>
<td>45.89</td>
<td>78.83</td>
</tr>
<tr>
<td>LIWC</td>
<td>57.14</td>
<td>38.13</td>
<td>45.74</td>
<td>78.20</td>
</tr>
<tr>
<td>Semantic lexicons</td>
<td>56.02</td>
<td>50.78</td>
<td>53.27</td>
<td>78.54</td>
</tr>
<tr>
<td>Spatial prepositions</td>
<td>41.67</td>
<td>3.47</td>
<td>6.40</td>
<td>75.57</td>
</tr>
<tr>
<td>Semantic distance</td>
<td>66.29</td>
<td>20.45</td>
<td>31.26</td>
<td>78.33</td>
</tr>
<tr>
<td>Classifer voting</td>
<td>66.84</td>
<td>67.76</td>
<td>67.30</td>
<td>84.13</td>
</tr>
<tr>
<td>Feature fusion</td>
<td>63.92</td>
<td>60.49</td>
<td>62.15</td>
<td>82.25</td>
</tr>
<tr>
<td>Meta learner</td>
<td>73.61</td>
<td>59.45</td>
<td>65.77</td>
<td>85.09</td>
</tr>
</tbody>
</table>

Table 5: Micro-averaged sentence-level results (%) on the test set (20% of all products). Maximum value in each column is boldfaced.

<table>
<thead>
<tr>
<th>Feature Type</th>
<th>Prec.</th>
<th>Rec.</th>
<th>F-score</th>
<th>Accu.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majority</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>76.13</td>
</tr>
<tr>
<td>Word unigrams</td>
<td>71.82</td>
<td>58.09</td>
<td>64.23</td>
<td>85.92</td>
</tr>
<tr>
<td>Meta learner</td>
<td>76.92</td>
<td>58.82</td>
<td>66.67</td>
<td>87.20</td>
</tr>
</tbody>
</table>

comparison, the table also shows the performance of the word unigram classifier, as well as a majority class baseline that labels every sentence as “non-usage.” As before, the meta-learner significantly improves over the unigram classifier,\(^5\) and also over the majority class baseline.\(^6\)

We also report the performance of the meta-learner classifier on individual products in Table 6. Across all the products, vinegar appears to have the highest F-score. This can be partly explained by the high inter-annotator agreement: the same product had the highest three-way agreement in the manual annotations, as shown in Table 3, likely an indication of a less difficult dataset.

### 6 Additional Analyses

To gain further insights, we perform several additional analyses, to determine: the role played by different features; the relation between classifier performance and amount of training data; the role of in-domain vs. cross-domain classification; and finally the types of errors produced by the system.

#### 6.1 Feature Importance Ranking

Table 7 shows the top features (ranked by their Gini importance (Breiman et al., 1984)) for three prominent individual feature-based classifiers — viz. word unigrams, category words, and concreteness — and the meta-learner. Note that top-ranking words include product properties (smell), secondary objects on which the product was used (clothes), how the product was used (day, daily, drink, water), usage verbs (use), prepositions and conjunctions (and, for, with), pronouns (i, it, this), and articles (a, the). For the meta learner, lexical features (character trigrams and word unigrams) and embedding features (Word2vec) are among the top-ranked feature classes.

#### 6.2 Learning Curve

Next, we experiment with varying the size of the training data to understand the learning curve. We gradually increased the amount of training data from 10% to 80%, in steps of 5%; and evaluated on the full test data. Figure 1 shows the variation of F-score achieved by the meta-learner as

![Learning Curve](image-url)

Figure 1: Learning curve using micro-averaged sentence-level results (%) per product using the meta learner.
Table 7: Feature importance ranking for four feature types. We show ten top-ranked features along with their importance scores. For the meta-learner, we show the ranking over the subset of seven feature sets used in this classifier.

<table>
<thead>
<tr>
<th>Feature Type</th>
<th>Prec.</th>
<th>Rec.</th>
<th>F-score</th>
<th>Accu.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>76.39</td>
</tr>
<tr>
<td>Word unigrams</td>
<td>69.15</td>
<td>35.20</td>
<td>46.65</td>
<td>80.99</td>
</tr>
<tr>
<td>Meta-learner</td>
<td>70.62</td>
<td>38.43</td>
<td>49.77</td>
<td>81.69</td>
</tr>
</tbody>
</table>

Table 8: Cross-domain classification: Micro-averaged sentence-level results (%), where test set is an individual product, and training set is four other products. Maximum value in each column is boldfaced.

<table>
<thead>
<tr>
<th>Feature Type</th>
<th>Prec.</th>
<th>Rec.</th>
<th>F-score</th>
<th>Accu.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>78.24</td>
</tr>
<tr>
<td>Word unigrams</td>
<td>74.19</td>
<td>50.74</td>
<td>60.26</td>
<td>85.44</td>
</tr>
<tr>
<td>Meta-learner</td>
<td>76.53</td>
<td>55.15</td>
<td>64.10</td>
<td>86.56</td>
</tr>
</tbody>
</table>

the training data is increased, smoothed over three consecutive data points. The test performance was the highest when trained on 60% of training data and then decreased gradually, which suggests that the system might not benefit from additional training data.

6.3 The Role of In-Domain Data

To understand the role played by in-domain data, we further experiment with two different configurations of training and test sets.

In one configuration, we train on four products, and test on the remaining product (cross-domain training). As can be seen from Table 8, this results in lower F-scores than Table 5. This suggests that identifying usage expressions of a product is intimately related to the identity of the product, echoing the findings by Govers and Schoormans (2005).

In the second configuration, we train on 80% of a product, and test on 20% of the same product (in-domain training). The results, averaged over the five products, are shown in Table 9. Note that the F-score values are much improved compared to the previous configuration, and are comparable to the results shown in Table 5. This suggests that when storage/memory might be a concern, we could simply use training data from within the do-

Table 9: In-domain classification: Micro-averaged sentence-level results (%), where test set is 20% of an individual product, and training set is 80% of the same product. Maximum value in each column is boldfaced.

<table>
<thead>
<tr>
<th>Feature Type</th>
<th>Prec.</th>
<th>Rec.</th>
<th>F-score</th>
<th>Accu.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>78.24</td>
</tr>
<tr>
<td>Word unigrams</td>
<td>74.19</td>
<td>50.74</td>
<td>60.26</td>
<td>85.44</td>
</tr>
<tr>
<td>Meta-learner</td>
<td>76.53</td>
<td>55.15</td>
<td>64.10</td>
<td>86.56</td>
</tr>
</tbody>
</table>

main to achieve comparable performance. This strategy also results in a faster training time and a smaller model, similar to the findings in (Bucilă et al., 2006).

6.4 Error Analysis

Finally, we also conducted a manual inspection of two broad categories of errors: false positives, i.e. “not usage” sentences marked as “usage” \((n = 25)\), and false negatives, i.e. “usage” sentences marked as “not usage” \((n = 56)\). This analysis revealed the following sub-categories for the false positives:

- **Number expressions**: Seven instances (29.17%) of errors can be attributed to numeric expressions occurring within sentences (“two years”, “3am”, “third bottle”, etc.).

- **Erroneous gold labels**: Six instances (25%) were actually correctly labeled as “usage” by the system, whereas the gold label was wrong ("I really love the smell of fresh laundry, and the smell of Downy.").

- **Shortcomings**: Six examples (25%) talk about actual or perceived shortcoming(s) of a product. “Olive oil used for healthy properties doesn’t keep well in plastic.[sic]"
• **Others**: Five instances (20.83%) were not captured by the above categories: “I used to drink a small shot each day, but haven’t for a while.”

False negatives have the following sub-categories:

• **Positive adjectives and adverbs**: 21 instances (37.5%) can be attributed to positive adjectives (“good”, “great”, “excellent”), and/or positive adverbs (“really”, “impressively”, “well”). “It smells amazing and lasts forever.”

• **Use-related verb in primary clause**: Eleven examples (19.64%) contain a use-related verb (“use”, “help”, “need”) in the primary clause: “I use this to eat, not to cook with.”

• **Erroneous gold labels**: Nine instances (16.07%) are actually correctly labeled as “not usage” by the system, but the gold label was wrong (“When I have to hang dry clothes, they get this horrible egg water odor.”).

• **Non-traditional usage**: There are three instances (5.36%) that talk about non-traditional or innovative usage of a product: “I have since made small sachet bags for my closets, car and as gifts.”

• **Others**: Twelve instances (21.43%) were not captured by the above categories: “I actually saw results after the first use.”

7 Conclusion

In this paper, we introduced the task of identifying usage expression sentences in consumer product reviews. A dataset comprising more than 3,000 annotated sentences was created from reviews of five products. We also trained a binary classifier to identify sentences that talk about the usage of a product. Extensive feature tuning and fusion experiments resulted in performance values comparable to the inter-annotator agreement. Detailed feature ranking, error analysis, and per-product performance numbers have been reported. Directions for future research include: experiments on a larger dataset of reviews with more diverse product types, expanding to other genres of reviews such as product blogs, and identifying types of usage expressions (how, where, why, and non-traditional uses). The work can also be extended to model the “personality” of a product with the “personality” of users – perhaps measured by the average personality of all people using the target product.

The annotated dataset is publicly available for research use from [http://lit.eecs.umich.edu/downloads.html](http://lit.eecs.umich.edu/downloads.html).

Acknowledgments

We thank Charles Welch, Aparna Garimella, Erin Donahue, Katie Cox, and Michelle Huang for their help with the annotations; Srayan Datta and Soumik Mandal for many helpful discussions and ideas. This material is based in part upon work supported by the Michigan Institute for Data Science. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the Michigan Institute for Data Science.

References


Xuan-Hieu Phan. 2006. CRFTagger: CRF English POS Tagger.


Chen-Tse Tsai, Gourab Kundu, and Dan Roth. 2013. Concept-Based Analysis of Scientific Literature. In *Proceedings of the 22nd ACM international conference on Conference on information &#38; knowledge management*, CIKM ’13, pages 1733–1738, New York, NY, USA. ACM.


