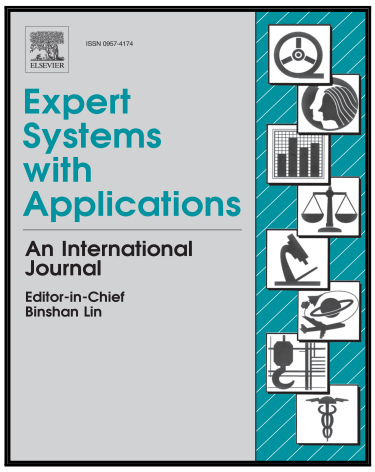

Automated defect discovery for dishwasher appliances from online consumer reviews

Darren Law , Richard Gruss , Alan S. Abrahams

PII: S0957-4174(16)30467-5
DOI: 10.1016/j.eswa.2016.08.069
Reference: ESWA 10862

To appear in: *Expert Systems With Applications*

Received date: 24 May 2016
Revised date: 29 August 2016
Accepted date: 30 August 2016

Please cite this article as: Darren Law , Richard Gruss , Alan S. Abrahams , Automated defect discovery for dishwasher appliances from online consumer reviews, *Expert Systems With Applications* (2016), doi: 10.1016/j.eswa.2016.08.069

Highlights

- Online dishwasher reviews contain many postings relating to dishwasher defects.
- We assess the effectiveness of sentiment analysis for dishwasher defect discovery.
- We propose new smoke term dictionaries for enhancing dishwasher defect discovery.
- Smoke terms deliver comparable performance to the best sentiment-based technique.
- Dishwasher smoke terms are distinct from sentiment terms, and mainly non-emotive.

Automated defect discovery for dishwasher appliances from online consumer reviews

Darren Law, Richard Gruss [a] *, Alan S. Abrahams [a]

[a] Department of Business Information Technology, Pamplin College of Business, Virginia Tech, 1007 Pamplin Hall, Blacksburg, VA 24061, United States

* Corresponding Author. Telephone: +1.540.449.6430, Fax: +1.540.231.3752

Email addresses: dlaw@vt.edu (D.S. Law), rgruss@vt.edu (R. Gruss), abra@vt.edu (A.S. Abrahams)

Abstract

Product defects can have a devastating impact on a firm's sales and reputation, especially in the era of social media. The early detection of defects could not only protect consumers from financial losses, but could also mitigate financial damage to the manufacturer. Previous work in automated defect discovery has had success in the automotive, consumer electronics, and toy industries, but so far there has been no application to home appliances. In this study, we extend the text analytic framework conceived in earlier work to the discovery of underperformance in large home appliances, specifically dishwashers. We find that generic cross-domain sentiment techniques can be strongly complemented by domain-specific "smoke" and "sparkle" term lists that are highly correlated with potential defects. These findings can be highly beneficial to improving dishwasher appliance quality management methods.

Keywords

Defect discovery, Text mining, Quality management

1. Introduction

Because dishwashers combine electricity, heating elements, electronic circuitry, and water, manufacturing defects can lead to serious hazards. The Consumer Product Safety Commission (CPSC), the federal agency tasked with investigating the safety of mass-produced appliances (CPSC, 2015) issued recalls for over 3 million dishwashers since 1999 due to fire hazards (CPSC, 2016) . Safety defects are a concern for manufacturers because they can result in significant costs and a loss of brand good will. In 2010, Maytag set aside $75 million to cover the costs of a recall of 1.7 million dishwashers (CBS News, 2015): a dozen consumers had reported that electrical failures in the dishwasher heating element led to fires and damage.

As damaging as safety defects are, they are fortunately rare. Far more common are performance defects, where products fail to meet consumer expectations of quality. These defects are less visible in the popular press, but are increasingly discussed in social media such as forums, online reviews, and social media.

Consumers are beginning to consult online reviews more frequently before making purchasing decisions. According to Nielsen, 70% of consumers trust online reviews (Nielsen, 2015), citing them as their second most trusted source after immediate friends and family. Two recent surveys of more than 2,000 American adults had the following findings: 1) 81% of Internet users have done online research on a product at least once, 2) 20% read online postings every day, 3) 73-87% report that reviews had a significant influence on their purchase, and 4) users are willing to pay from 20-90% more for a 5-star-rated product than a 4-star-rated product (Horrigan, 2008; Lipsman, 2007). These figures indicate that online review websites and social media channels can significantly affect user perceptions, and therefore, can directly influence sales. Electronic word of mouth (eWOM) consisting of mainly negative opinions

about a product has been shown to be related to reduced sales of that product (Kim, Wang, Maslowska, & Malthouse, 2016).

Because of their detail and candor about product performance, online reviews could be a valuable resource for the detection of design and manufacturing defects. The large volume, however, presents serious challenges. Information is scattered across the thousands of reviews on multiple sites, and it is difficult to synthesize it all into actionable business intelligence. An efficient automated system of isolating product defects in online reviews is needed.

In this study, we create and evaluate a system that identifies potential dishwasher performance defects in online reviews by leveraging text analytic techniques. A text analytic system for large scale surveillance of dishwasher appliances from online reviews could be helpful to a number of stakeholder types. These reviews contain valuable information and could provide critical timely feedback to companies (Abrahams, Jiao, Wang, & Fan, 2012). Particularly, manufacturers seeking to improve their manufacturing processes to better serve consumers would find this valuable. This research could also benefit property owners who want to install the best dishwashers, or engineers seeking to develop innovative new appliances to counter common complaint themes.

The rest of this paper is structured as follows: in Section 2 we describe the background to this study, examine prior work relevant to our research, and identify critical gaps; in Section 3 we present research questions and contributions; in Section 4 we describe our methodology for collecting and analyzing online reviews, including two experiments that test our dishwasher defect discovery techniques; in Section 5 we discuss our findings, and finally, in Section 6, we suggest areas for future research.

2. Background and Related Work

In this section, we provide some background on the nature of online reviews. We then examine related work in the field of text classification, sentiment analysis, and quality assurance. We discuss how these related works contribute to the idea of automated online review analysis, and review the shortcomings of each type of work.

2.1 Online Reviews

For the purposes of this study, we restrict the definition of 'online reviews' to websites that provide large volumes of product-specific, public, postings. Examples include: Amazon.com, BestBuy, and Consumer Reports. Websites such as Facebook and Twitter that do not provide large volumes of product-specific and public postings are excluded from consideration because their user-posted product-specific content is not readily accessible, due to this content either being private or not product-targeted. The scattered nature of the many reviews on Facebook and Twitter presents a significant challenge in finding, consolidating, and garnering meaningful insights from them. Facebook, for example, has company and product pages, controlled by moderators who not only market products, but also collect

information regarding product complaints ("Facebook. GE Appliances Company ", 2016). However, identifying the specific product being addressed in Facebook postings can be a significant challenge. Our analysis is restricted to public, product-specific online reviews, such as those found on retailer websites.

2.2 Text Classification

The term "text classification" describes a process in which documents are classified into predefined categories based on textual content. Text classification has been applied to several practical tasks such as news article classification (Aggarwal & Zhai, 2012; Lang, 1995), spam detection (Chakrabarti, Dom, Agrawal, & Raghavan, 1997), and document indexing (Carvalho & Cohen, 2005; W. W. Cohen, 1996; Lewis & Knowles, 1997; Sahami, Dumais, Heckerman, & Horvitz, 1998).

For some applications, textual information needs to be supplemented with other document metadata to achieve acceptable accuracy. Document classifiers have been trained with other factors aside from post content, including author information, time of event, and other meta-attributes (Sriram, Fuhry, Demir, Ferhatosmanoglu, & Demirbas, 2010).

Various researchers have acknowledged the many important applications of text classification, and improving accuracy has been a dominant concern in the field. Especially critical to classifier performance is the collection of good quality, labeled, training sets. Large numbers of documents are not necessarily required to train good text classifiers: for example, a modified version of the Naïve Bayes classifier requires fewer inputs to train the classification system (Kamruzzaman, Haider, & Hasan, 2010).

In the appliance industry, there is a need for categorizing reviews with defects. With valuable feedback from a text classification system, companies could more effectively use data for quality management.

2.3 Sentiment Analysis

Sentiment Analysis is a widely used data mining tool to analyze opinions and discover how consumers feel. For example, the Harvard General Inquirer sentiment analysis tool takes a pre-set list of positive and negative words and analyzes the polarity of documents based on the prevalence of words from each category.

However, accuracy continues to be a concern in sentiment analysis. The OpinionFinder system seeks to address this problem by analyzing polarity in context, rather than simply through single-word triggers. For example, OpinionFinder assigns a negative polarity to the phrase "not good", as the modifier "not" adjusts the polarity of the positive word "good" (Abrahams, et al., 2012; Riloff & Wiebe, 2003; Wiebe & Riloff, 2005; Wilson, Wiebe, & Hoffmann, 2005).

In addition to evaluating the polarity of product reviews, research has expanded to areas such as opinions of trending topics in social media, and the variance of sentiment by demographics and location. For example, a corpus of 300,000 twitter posts was collected to create a sentiment classification system to

automate the process of evaluating polarity of Twitter posts (Pak & Paroubek, 2010). Godbole, Srinivasaiah, and Skiena (2007) created "sentiment maps" to show how sentiment varies based on the demographic group, news source, or geographic location .

In our research, we explore alternatives to sentiment analyses, since prior research indicates defects may not be associated with negative sentiment, but instead may be pinpointed using domain-specific "smoke" words (Abrahams, et al., 2012). "Smoke" words are words highly prevalent in defects, relative to non-defects, and may include non-emotive (non-sentiment) words. By creating smoke word lists consisting of words and phrases specific to the industry of interest – including both emotive (sentiment) and non-emotive words – we hope to determine if defects can be discovered more accurately.

2.4 Quality Assurance in Dishwasher Appliance Manufacturing

Quality assurance techniques are used both on the supply- and demand-side. On the *supply-side*, the primary technique of acceptance sampling is Statistical Process Control (SPC). This analytic measurement system, focused on maintaining process performance, is used to ensure that products meet the manufacturer's specification standards. SPC also aims to find which factors contribute most to optimal performance, such as improved quality and lower costs (Woodall & Montgomery, 1999). Another statistical approach is Six Sigma, which is implemented by many firms to minimize defects. Aimed at producing only 3 - 4 defects per 1,000,000 products, Six Sigma measures, analyzes, improves, and controls processes. Any deviation in specification indicates quality reduction (Crosby, 1980). Appliance manufacturers are actively employing Six Sigma for quality control. Jack Welch, CEO of GE from 1981-2001 implemented many aspects of Six Sigma to boost profitability and performance (Hahn, Hill, Hoerl, & Zinkgraf, 1999). There are also several lean manufacturing methods designed to increase efficiency and reduce manufacturing process steps. Taiichi Ohno's famous Toyota Production System creates value-added steps designed to indirectly reduce defects (Nelson-Peterson & Leppa, 2007; Ohno, 1988).

On the *demand-side*, firms monitor feedback from consumer reports, surveys, and direct consumer responses and complaints (via customer support call centers or e-mail). The manufacturer uses the consumer's perceptions of quality to adjust standards of production. Ideal or preferred attributes of a product can also influence manufacturing techniques (Abbott, 1956; Garvin & Quality, 1984; Griliches, 1971; Johnson, 1971; Lancaster, 1971; Leffler, 1982). Demand-side feedback provides valuable information to producers by creating new goals for continuous improvement.

2.5 Defect Discovery from Social Media

Social media surveillance, text classification, and sentiment analysis have been used successfully in previous work on defect discovery. Prior researchers have developed text analytic frameworks for defect discovery and applied these methods to defect discovery in the automotive and consumer

electronics industries (Abrahams, Fan, Wang, Zhang, & Jiao, 2015; Abrahams, Jiao, Fan, Wang, & Zhang, 2013; Abrahams, et al., 2012) and the toy industry (Winkler, Abrahams, Gruss, & Ehsani, 2016). The classifiers built for automotive-, consumer electronic-, and toy-defect discovery, are inappropriate for dishwasher defect discovery due to the different components of the product, and different failure modes. Further, prior defect-discovery classifiers were limited to simple unigram (single word) term lists, and numeric threshold-based classification and logistic regression.

The current study extends term lists to bi-and tri-grams (two- and three-word phrases), and applies and evaluates additional classifiers, including neural networks and decision trees. Finally, prior defect-discovery classifiers have overlooked the importance of the product-category-specific *performance expectations*. In this paper, we define and assess a new class of domain-specific 'sparkle' words, which capture both consumer satisfaction and the compliance of the product with its specification or intended usage.

In this paper, we develop, apply, and assess novel industry-specific smoke-word and content-tagging constructs for defect discovery in the dishwasher industry, to provide business intelligence to dishwasher manufacturers for continuous product improvement. Developing new constructs that address the particular components and failure modes for dishwashers, and assessing the validity and reliability of these new constructs, is novel and important, and addressed in this work. In addition, we expand the assessment of classifiers to include decision trees and neural networks, rather than simply numeric thresholds and logistic regression.

3. Research questions and contributions

In this paper, we address three research questions. First, how prevalent are defects in online reviews in the dishwasher industry and how can such defects be categorized? Second, what text analytic methods (sentiment analysis, significant terms) will provide the best performance in defect discovery from online dishwasher reviews? Third, which terms are of highest significance in detecting defects in online reviews in the dishwasher industry?

We present major contributions to the dishwasher industry from this research. Automated defect detection system of online reviews offers benefits to companies in areas such as quality management and product improvement. We demonstrate how various methods of sentiment analysis can be applied to accurately detect defects in online reviews. Finally, we create multiple new domain-specific dictionaries that will allow companies in the future to better find and apply valuable defect-related information in the dishwasher industry. These dictionaries include smoke-term dictionaries specific to the dishwasher industry, as well as a new domain-specific term list, which we dub 'sparkle' terms, that capture the

meeting or exceeding of the customer's product-category specific expectations, and satisfactory delivery to product specification.

Substantial effort is required to gather dishwasher-specific reviews, develop reliable tagging constructs, manually tag large volumes of dishwasher reviews for the existence of different defect types and affected components, and assess the developed constructs. This manuscript develops, describes, and tests the validity, reliability, and generalizability of novel domain-specific constructs. These novel, re-usable assets (smoke terms, sparkle terms, content dictionaries and scores, and procedure) are valuable for future researchers, who can employ them on future studies with the assurance that they perform well. Prior studies do not develop constructs specific to the dishwasher industry, nor demonstrate that the constructs are valid, reliable, and generalizable to dishwashers. Both the development of the industry-specific dishwasher constructs and their assessment in the dishwasher industry are contributions to the literature.

4. Methodology

We obtained 11,024 online product reviews from two sources: Amazon.com and Bestbuy.com. The Amazon reviews dated from May 1996 to July 2014 (McAuley, Pandey, & Leskovec, 2015) and were filtered to include only the dishwashers category. The Best Buy reviews dated from January 2010 to September 2015 and were also filtered to include only dishwashers.

4.1 Sample Statistics

From the total reviews gathered from Amazon and Best Buy, the statistical distribution of reviews is shown in Table 1:

Table 1: Number of Star Reviews in Each Dataset

Review Type	Amazon	Best Buy
1 Star Review	1,742	358
2 Star Review	428	164
3 Star Review	277	379
4 Star Review	614	1880
5 Star Review	1,559	3,622
Grand Total	**4,620**	**6,404**

Within the Amazon reviews, there were a total of 730,124 words, and an average number of 175 words per review. Within the Best Buy reviews, there were a total of 306,421 words, and an average of 48 words per review. The total number of reviews in the Amazon dataset for the dishwasher product category was 4,620, and the total number of reviews in the Best Buy dataset for the dishwasher product category was 6,404.

4.2 Data Coding

We assembled a labeled training set classifying reviews into the following target categories: "Performance Defect," "Safety Defect", and "No Defect" (Abrahams, et al., 2012). These are each described below:

1. "Performance Defects" are reviews that indicate a non-serious product malfunction unlikely to cause injury. These defects are typically related to consumer satisfaction, such as how well the dishwasher cleans, or how long the dishwasher lasts before there is a problem. In this actual example, a consumer complains about the cleaning features of a dishwasher:
"[Brand A] is the worst dishwasher I've ever owned, I have to wash off every spot before putting them in the dishwasher. Most of the time it is easier just to wash my dishes by hand. I have owned dishwashers for the past 45 years and this is the worst ever."

2.) "Safety Defects" are reviews that indicate a serious product malfunction likely to cause serious injury or property damage. Examples include dishwasher fires which can cause major property damage and consumer casualty. The following is an actual example of a safety defect from a consumer review.
"Dishwasher caught on fire.. walked into house and smoke was billowing everywhere!"

3.) "No Defects" are reviews that may contain irrelevant information, or do not indicate a product failure. Examples include positive product reviews, advertisements, or general comments.
"I haven't used the dishwasher yet. Fit great & looks amazing. Bought it for a "flip" house. No room for a full size dishwasher. I am sure it will help us to get top dollar for the house."

Additional fields such as dishwasher component category and injury severity were collected for supplementary data. These are described in Appendix A.

4.3 Data Processing

A random sample of 4,168 Amazon reviews was chosen. Nine undergraduate business students were asked to tag at least 500 reviews each, using the tagging protocol in Appendix A, which develops a number of dishwasher-specific tagging constructs. Three students chose not to participate, and one student tagged only 180 reviews. The lead researcher also tagged 500 reviews. In all cases, reviews were randomly selected from the random set, and taggers were blind to the selections of other taggers. Taggers tagged a total of 3,216 reviews, though due to the random selection presented to

each tagger, only 2,309 unique reviews were tagged: 1,576 reviews were tagged by more than one tagger, and the remaining 733 reviews were tagged by only a single tagger. To establish inter-rater reliability, the tags of the lead researcher were compared to the tags of the other taggers. Due to random review presentation, 354 of the lead researchers' 500 tagged reviews overlapped with other taggers. Inter-rater agreement statistics for each attribute tagged are shown in Table 2, which indicates Cohen's κ (J. Cohen, 1968) for each attribute, as well as the categorical interpretation of the extent of agreement indicated by Cohen's κ for each attribute, per Landis and Koch (Landis & Koch, 1977), and Fleiss et al. (Fleiss, Levin, & Paik, 2013). Table 2 indicates that all dishwasher-specific constructs developed have acceptable reliability.

Table 2: Inter-Rater Reliability, by Attribute, for Dishwasher-specific Constructs in Training Set.

Attribute	Percent Agreement	N Agreements	N Disagreements	Cohen's κ	Extent of Agreement	
					(Landis & Koch, 1977)	(Fleiss, et al., 2013)
Defect Severity	0.873	309	45	0.809	Almost perfect	Excellent
Injury Timing	0.992	351	3	0.989	Almost perfect	Excellent
Defective Component	0.760	269	85	0.726	Substantial	Fair to good
Complaint Type	0.655	232	122	0.606	Substantial	Fair to good

These tagged reviews were used to create a *training set*. As some reviews were tagged by multiple taggers, the training set was conflict-resolved using a "Majority Conservative" conflict-resolution strategy: the majority vote of the taggers was used as the chosen classification; in the case of voting ties, the most conservative category was chosen (safety defect over performance defect, and performance defect over no defect). The conflict-resolved training set was necessary because reviews were distributed randomly to taggers, to reduce bias; many of the reviews were tagged multiple times by different taggers. Then, using the Correlation Coefficient (CC) metric (Fan, Gordon, & Pathak, 2005), we analyzed which unigrams (one word), bigrams (two word clusters), and trigrams (three word clusters) were most prevalent in reviews indicative of defects, compared to those containing no defects. These terms (unigrams, bigrams, and trigrams) were organized into separate "smoke" term lists. In the creation of these "smoke" term lists, brand names, including retailer and manufacturer names, were excluded, to reduce popularity bias (top selling models and manufacturers appear more frequently in defects, due to larger sales volumes).

4.3.1 Experiment 1: Validating the Smoke Word Lists

For Experiment 1, we created a combined set of 8,251 Amazon and Best Buy reviews to test each smoke word dictionary (a total of 6,403 from Best Buy, and a total of 1,848 from Amazon.com). This set of reviews was an unseen *holdout* sample, meaning that the reviews from the initial training set were not included, to reduce training bias.

The smoke term lists that were tested included: Unigram Smoke Word Dictionary, Bigram Smoke Word Dictionary, and the Trigram Smoke Word dictionary. Reviews in the holdout sample were scored by accumulating a score based on how frequently the smoke unigrams, bigrams, and trigrams were used in the reviews, and the relative weight (importance) of each unigram, bigram, or trigram as assessed using the CC metric.

To compare the relative performance of the smoke word lists to conventional sentiment analysis, we also scored the sets using three sentiment dictionaries: the Harvard General Inquirer's Negative word list (Edward & Stone, 1975), AFINN (F. Å. Nielsen, 2011), and ANEW (M. M. Bradley & Lang, 1999). Again, each review was scored by accumulating a total score based on the frequency and relative weight of each word occurrence.

After applying each smoke word dictionary and sentiment analysis method to the unseen holdout data set, we sorted the reviews in descending order by the score for each method, and analyzed the top and bottom 200 reviews for each method. As many of bottom 200 reviews shared zero scores, we shuffled and randomly selected these reviews to reduce bias. We then compiled each of these top and bottom 200 reviews to create a *validation set* of 2,400 reviews. These 2,400 reviews were again randomly distributed to the same team of undergraduate business students. Each student was again presented with 500 reviews in random order, and again given the protocol shown in Appendix A. To reduce bias, the team of taggers was shown the review text only, with no indication of any review scores.

Some reviews in the validation set were duplicates, appearing in the top 200 or bottom 200 for multiple methods. In total, there were 1,475 unique reviews. 791 reviews were tagged by multiple taggers, and 684 reviews were tagged by a single tagger, for a total of 2,835 validation tags. Tagging by a single tagger was accepted as sufficient tagger reliability was established during tagging of the training set. Once all of the reviews were tagged, they were again conflict resolved to come to final decisions. The purpose of this second tagging session was to assess the performance of the various smoke word and sentiment analysis methods in determining defects in the dishwasher appliance industry. We discuss our results and findings in Section 5.1.

To determine the association between consumer's star rating and defect existence, we found the total number of reviews per star-rating – see Table 3. Table 3 indicates that star-rating cannot be used

as a singular mechanism for detecting defects: although the proportion of defects is higher in low-star reviews than in high-star reviews, it is clear that many four- and five-star reviews still discuss defects.

Table 3: Number of "No Defect" and "Performance Defect" per Review Star-Rating

Star Rating	No Defect	Performance Defect	Grand Total
1	49	863	912
2	17	184	201
3	41	93	134
4	243	66	309
5	691	59	750
Grand Total	**1,041**	**1,265**	**2,306**

4.3.2 Experiment 2: Training and Validating Classifiers

To create a tabular feature set for input into classifiers, we employed the smoke term lists from Experiment 1, and the tagging results from the training set, and developed domain-specific term lists, to capture performance issues, safety issues, specific component mentions, and specific outcomes mentions. Further, we defined a set of domain-specific 'sparkle' words, which capture the consumer's product-category-specific expectations for dishwashers being met or exceeded, or, compliance with the product category's design specification. In each case, the domain-specific term lists were created by contrasting words in reviews in each training-set tag category (see Appendix A) to words in reviews outside those tag categories, again using the CC metric, followed by manual filtering by a member of the research team.

Table 4 shows the domain-specific term lists developed, and provides an itemization of the top 20 terms from each list. For each term list, we accumulated a score for each review if a term (word or phrase) from the list appeared in the review. In addition to the term-list-specific scores, we appended traditional sentiment scores (gross number of AFINN negative words, and gross number of Harvard General Inquirer negative words), to the feature set.

We used the feature set of review scores from domain-specific term lists (Table 4) and conventional sentiment scores for each review, as input to three popular classifier types in *JMP Pro 12*: Nominal Logistic Regression, Decision Tree Partition, and Neural Network. 10-fold cross validation was performed for Decision Trees and Neural Nets. As JMP does not support 10-fold cross-validation for nominal logistic regression, a random 90% training 10% validation was performed for nominal logistic regression.

Table 4: Top 20 terms from each dishwasher domain-specific term list

List Type	List Name	Top 20 Terms
Smoke Words	Performance Issues - Unigrams	service, repair, warranty, replaced, after, customer, again, call, new, called, replace, months, worst, will, told, years, fix, another, year, stopped
	Performance Issues - Bigrams	a new, customer service, had to, I called, would not, to replace, the control, the same, piece of, not buy, the worst, replace the, does not, the warranty, the part, the motor, control panel, do not, the wheels, replaced the
	Performance Issues - Trigrams	does not clean, the top rack, had to be, to replace the, the control panel, is the worst, have had to, I was told, this piece of, a waste of, We have had, had to replace, have had the, we have had, will never buy, do not buy, not buy this, an extended warranty, to be replaced, the worst dishwasher
	Safety Issues	burning, smelling, trauma, recalling, susceptible, damage, braker, hazardous, wisp, charring, burnt, melting, scared, threatening, skin, fingertips, wires, caught, burn, ruins
Outcome-specific words	Cleanliness Issues	clean, dishes, film, dirty, worse, leaves, agents, rinse, wash, cleaned, bakes, food, dirtier, detergents, pitiful, particles, glasses, detergent, rinsing, scum
	Noise Issues	noise, loud, whistling, grinding, decibels, noises, chop, loudly, squeaking, decibel, percussive, emanating, rubs, moaning, chronic, rhythmically, sound, whistle, sounded, louder
	Leak Issues	leak, leaking, leaked, leaks, flooded, puddle, seals, reglued, floor, gasket, towels, dumped, slipped, flooding, hardwoods, mouldy, caulked, carpeted, liters, linoleum
	Smell Issues	odor, moldy, soggy, underwear, reeks, stinky, smell, stagnant, smells, smelling, slimy, disgusted, excessive, rotten, spoiled, burning, milk, smelled, clog, terrible
Component Words	Hose	separates, ply, hose, blocks, freezing, caulked, mop, blocked, supply, adaptors, fastened, gfci, backflow, sudsed, clamp, inlet, hoses, fitting, obstruct, route
	Latch	latch, reattached, reglue, latching, weld, stiffener, unglued, welding, springs, reattach, mechanism, reassemble
	Button	panel, control, touchpad, button, buttons, reboot, corrodes, bursting, touchy, resistor, capacitor, haywire, buttoms, illuminating, triggered, operable, diagnostic, function, touch, sensitive
	Spray Arm	arm, spray, spindle, spinning, warped, plumbers, whirring, agitator, grinds, sprayer
	Pump	pump, disconnection, grinding, motor, pumps, circulator, pumping
Sparkle Words	Unigrams	easy, love, perfect, works, great, small, fits, cleans, sparkling, best, happy, amazing, quiet, perfectly, recommend, pleased, excellent, highly, spotless, glad

5. Results and Evaluation

This section evaluates the performances of the smoke word dictionaries and sentiment analyses, and discusses our findings. Because there were not enough Safety Defects in the tagged set to find significant patterns in Safety Defects alone – only 11 of 2,321 reviews (less than half a percent) were Safety Defects – we focused on Performance Defects only. "Performance Defects" and "No Defect" were then compared in assessing the performance of the sentiment analyses and smoke word dictionaries in defect discovery.

5.1. Experiment 1 Results

The defects discovered by the sentiment analysis approaches and smoke word dictionaries in the validation set are shown in Table 5:

Table 5: Defects Discovered, per Dictionary

Row Labels	No Defect	Performance Defect	Grand Total
AFINN	**138**	**262**	**400**
Bottom 200	15	185	200
Top 200	123	77	200
ANEW	**284**	**116**	**400**
Bottom 200	181	19	200
Top 200	103	97	200
General Inquirer Negative	**176**	**224**	**400**
Bottom 200	120	80	200
Top 200	56	144	200
Smoke Unigrams	**232**	**168**	**400**
Bottom 200	194	6	200
Top 200	38	162	200
Smoke Bigrams	**215**	**185**	**400**
Bottom 200	193	7	200
Top 200	22	178	200
Smoke Trigrams	**203**	**197**	**400**
Bottom 200	187	13	200
Top 200	16	184	200
Grand Total	**1248**	**1152**	**2400**

For each scoring method, we ran a Chi-squared test to verify whether the top 200 scoring reviews contained significantly more defects than the bottom 200. Our findings are shown in Table 6. Note that AFINN has low, negative scores for negative sentiment, ANEW has low positive scores for negative sentiment and high positive scores for positive sentiment, General Inquirer (GI) Negative

has increasingly positive scores for reviews containing more negative words, and the manual smoke word dictionaries have high scores for reviews that are expected to be indicative of defects.

Table 6: Comparison of Top vs Bottom Scoring Reviews for each Scoring Method:

Scoring Method	Finding	p-value(Chi-Test)
AFINN	Bottom 200 have statistically more defects than Top 200	<0.001**
ANEW	Top 200 have statistically more defects than Bottom 200	<0.001**
GI Negative	Top 200 have statistically more defects than Bottom 200	<0.001**
Unigram	Top 200 have statistically more defects than Bottom 200	<0.001**
Bigram	Top 200 have statistically more defects than Bottom 200	<0.001**
Trigram	Top 200 have statistically more defects than Bottom 200	<0.001**

**Indicates statistical significance at the 99.9% confidence level

The results of the sentiment and smoke word dictionary analyses determined that the dictionaries could isolate review content indicative of defective performance. The ANEW technique, however, had more defects in reviews that were expected to be non-defects. Exploring this anomaly, we found that the most positive (highest scoring) ANEW words -- “Water”, “Dishes”, and “Clean” -- were found in many reviews containing defects. This method was highly subject to false-positives. This phenomenon is illustrated in Table 7, where the ANEW word valence score is summed across all performance defects in the top 200 ANEW set.

Table 7: Top 20 Words Non-Indicative of Performance Defects (ANEW)

Word	Score
water	4071.76
dishes	2870.66
clean	2508.81
like	2496.64
me	2176.2
rack	2166.5
unit	1878.24
machine	1725.51
time	1677.96
good	1404.36
first	1288.43
door	1118.34
quiet	1088.1
model	1079.2
old	893.7
sure	796.66
know	790.02
right	786.9
space	786.48
easy	781

Using another Chi-squared test, we tested which scoring methods were significantly better at finding defects than a random selection of reviews. In the training set, 1,259 out of 2,310 reviews were “Performance Defect”, and 1,043 out of 2,310 reviews were “No Defect”. This equated to a baseline rate of 55% for “Performance Defect” and 45% for “No Defect”, and consequently a rate of 110 defects per 200 reviews and 90 non-defects per 200 reviews, respectively. From this conversion, we compared each method's performance in the validation set to the baseline expected rate of defects from the training set. The results are shown in Table 8.

Table 8: Comparison of each Scoring Method to Baseline Accuracy:

Scoring Method	Finding	p-value (Chi-Test)
AFINN	Bottom 200 have statistically more defects than baseline.	<0.001**
ANEW	Top 200 do not have statistically more defects than baseline.	0.07
GI Negative	Top 200 have statistically more defects than baseline.	<0.001**
Unigram	Top 200 have statistically more defects than baseline.	<0.001**
Bigram	Top 200 have statistically more defects than baseline.	<0.001**
Trigram	Top 200 have statistically more defects than baseline.	<0.001**

Next, we compared scores for each smoke word dictionary and sentiment analysis approach to determine whether each approach had significantly different scores for Performance Defects vs. Non-Defects. To determine whether these findings were statistically significant, we ran student's T-tests. These results are shown in Table 9.

Table 9: Statistical difference in Means, for Defects vs. Non-Defect, for each Scoring Method:

Scoring Method	Finding	p-value (T-Test)	Mean Score for No Defects	Mean Score for Performance Defects
AFINN	Defects have lower score	<0.001**	24.51	2.05
ANEW	Defects have higher score	<0.001**	151.69	346.66
GI Negative	Defects have higher score	<0.001**	4.43	10.55
Unigram	Defects have higher score	<0.001**	8,892.18	85,626.22
Bigram	Defects have higher score	<0.001**	1,690.94	24,691.77
Trigram	Defects have higher score	<0.001**	51,511.60	333,647.56

Lastly, we analyzed the sentiment method AFINN – the best performing sentiment method – to determine which individual sentiment words contributed most strongly to defect discovery (i.e. contributed the highest sum of accumulated negative points, in actual defects). Table 10 shows the AFINN words contributing most to negative sentiment in actual defects.

Table 10: Top 20 Words Indicative of Performance Defects (AFINN)

Word	Score
problem	-206
bad	-159
no	-155
worst	-147
problems	-116
error	-102
terrible	-96
died	-72
horrible	-66
dirty	-62
disappointed	-60
wrong	-60
awful	-54
worse	-54
failed	-52
poor	-52
hate	-51
stuck	-46

Note that the AFINN sentiment analysis was not statistically better at finding defects than the trigram smoke word dictionary. The difference between these two approaches was 1 defect (Table 5), which was not statistically significant ($p<0.001$). All of the custom smoke word dictionaries also outperformed ANEW and the Harvard GI Negative, and detected more defects as the word cluster size increased: trigrams detected more than bigrams which detected more than unigrams. This outcome may occur due to the ability of word clusters to exclude false-positives.

To test the similarity of word lists, we found which words appeared in both the AFINN sentiment dictionary, and the top 20 unigrams. The two words that appeared in both lists were "worst" (13^{th} most-contributing word) and "stopped" (20^{th} most-contributing word). This indicates that sentiment analysis and smoke word analysis are complementary techniques, with negligible word overlap. It is evident that defects in the dishwasher industry are frequently indicated by non-emotive (non-sentiment) words, reinforcing the findings of prior work in other industries (Abrahams, et al., 2015; Abrahams, et al., 2012) that generic sentiment analysis should not be solely relied on for defect discovery from online reviews. We also found that the trigram smoke word approach found 168 unique defects, amongst its top-scoring 200 reviews, that were not discovered in the 200 bottom-scoring (most negative sentiment) reviews ranked by the best-performing sentiment approach, AFINN. The two approaches shared only 17 identical defects in

their sorted list of 200 reviews most likely to contain defects. Smoke-scoring therefore is highly-complementary to sentiment-analysis, for defect discovery in the dishwasher industry. Appendix B provides some examples of a few of the unique defects that were discovered by the Top 200 Trigram smoke approach, but not found by the Bottom 200 AFINN (most negative sentiment) approach.

5.2. Experiment 2 Results

Table 11 below captures the results of the training and validation using the three classifier approaches. For each classifier, in the top, unshaded portion of Table 11, the parameters for the three features with the most impact are shown **bolded**, to highlight their importance.

Interpreting the odds ratios (i.e. Exp(β)) for the logistic regression from Table 11, we see, for instance, that for each additional term (trigram) found in the review from the "Performance Issues – Trigrams" list, the odds of the review being a defect increases by 77% ($\text{Exp}(\beta) = 1.77$). Similarly, for each additional 'sparkle' term found in the review, the odds of the review being a defect decreases by 31% ($1 - \text{Exp}(\beta) = 1 - 0.69 = 0.31$).

Specific dependent-variable coefficients for neural networks are not shown, as they are not easily interpretable. Instead, "/" indicates that JMP's neural network profiler shows variations in the variable across its range have a strong impact, whereas "—" indicates variations in the variable have little impact.

Table 11 shows that 'sparkle' terms, tend to have a strong influence in all models, and smoke scores (unigram, bigram, and trigram), are strongly influential in both logistic regression and decision trees. While conventional sentiment approaches are moderately influential, they typically are less influential than the domain-specific sparkle and smoke words for dishwasher appliances. Finally, component-specific words and outcome-specific words typically have little to no influence, most likely as each specific sub-category accounts for only a small portion of defects.

For model comparison, Table 11 reports a number of model Performance Metrics, including Generalized R^2, Root Mean Squared Error (RMSE), Brier score, Area Under Curve (AUC), and R^2 on ten-fold cross-validation. **Bolded** font under Performance Metrics, in the bottom, shaded portion of Table 11, indicates the best classifier according to that metric. Generalized R^2 and RMSE are common metrics of general statistical model performance, though Brier score and AUC are more specifically appropriate measures of binary classifier performance (Hernández-Orallo, Flach, & Ferri, 2012). The various model Performance Metrics can be interpreted as follows:

- *Generalized R^2*, also known as Nagelkerke R^2 (Nagelkerke, 1991), is a performance metric for general regression models, and is based on a likelihood function. Generalized R^2 is scaled to a maximum value of 1, though high values (approaching 1) are difficult to achieve for binary logistic models. Neural networks and decision trees tied for best performance on this metric.
- For *RMSE*, models with RMSE approaching 0 are regarded as better models, with lower residual errors – i.e. lower difference between actual defect existence and the fitted probability for defect occurrence. Decision trees edged out neural networks slightly on this metric.
- The *Brier score* (Brier, 1950) is a proper score function that measures the accuracy of probabilistic predictions, and is applicable to tasks in which predictions must assign probabilities to a set of mutually exclusive discrete outcomes, such as Defect vs. Non-Defect status. The Brier score is effectively the mean squared error of the prediction, with values approaching 0 indicating the best models, and values approaching 1 indicating the worst models. Brier scores were similar for all three models, with the neural network classifier slighting edging out logistic regression for best performance.
- For *AUC* (A. P. Bradley, 1997; Hanley & McNeil, 1982), values of AUC approaching 1 indicate the best models, which seldom have false positives in the top-scored items, and seldom have false negatives in the bottom-scored items. Values of AUC approaching 0 indicate the worst models, which frequently have false positives in the top-scored items, and frequently have false negatives in the bottom-scored items. AUC was very high (i.e. close to 1) across all candidate models, indicating a very low incidence of false positives amongst top-scoring items, and a very low incidence of false negatives amongst bottom-scoring items. AUC was statistically significantly better ($p<0.01$) for the best performing classifier (decision tree; AUC=0.94) compared to the classifier with the lowest performance (logistic regression; AUC=0.92), though there is little practical difference between the three approaches (Δ AUC < 0.02).
- The final row of Table 11 shows *R^2 on ten-fold cross-validation*, and indicates that the decision tree and logistic regression models are impacted by over-training, as they perform less well on the validation sets, compared to the neural network model which performs well on both training sets and validation sets.

Table 11. Classifier Results

		Logistic Regression				Neural Net	Decision Tree (Partition)		
Independent Variable		**LogWorth**	**β**	**Exp(β)**	**LogWorth FDR**	**Profiler**	**# Splits**	**G^2**	**Portion**
Smoke Words	Performance Issues – Unigrams	**7.1**	**0.08****	**1.08**	**6.3**	—	**6**	**179.9**	**0.10**
	Performance Issues – Bigrams	4.2	0.14**	1.15	3.8	—	**4**	**841.3**	**0.46**
	Performance Issues – Trigrams	**20.8**	**0.57****	**1.77**	**19.9**	—	7	170.4	0.09
	Safety Issues	0.0	0.00	1.00	0.0	—	2	11.6	0.01
Outcome Words	Cleanliness Issues	0.9	0.03	1.03	0.7	—	6	65.1	0.04
	Noise Issues	0.5	-0.10	0.91	0.4	—	0	0	0
	Leak Issues	1.3	0.20	1.22	1.0	—	0	0	0
	Smell Issues	1.0	0.26	1.30	0.8	—	0	0	0
Component Words	Hose	1.4	-0.12	0.88	1.0	/	2	14.5	0.01
	Latch	0.1	-0.09	0.91	0.1	—	0	0	0
	Button	0.4	-0.06	0.94	0.3	—	0	0	0
	Spray Arm	0.1	0.01	1.02	0.1	—	0	0	0
	Pump	1.2	0.36	1.44	0.9	—	1	5.6	0

Sparkle Words	Unigrams	**60.0**	**-0.36****	**0.69**	**58.7**	**/**	**9**	**487.2**	**0.27**
Sentiment Words	AFINN Negative	4.9	0.08**	1.09	4.4	—	0	0	0
	General Inquirer	6.2	-0.10**	0.90	5.6	/	2	20.8	0.01
Performance Metrics	Generalized R^2	0.62				**0.70**	**0.70**		
	RMSE	0.33				0.31	**0.30**		
	Brier score	0.094				**0.092**	0.107		
	AUC	0.921				0.938	**0.941**		
	R^2 on ten-fold cross-validation†	0.52				**0.71**	0.54		

LogWorth FDR = Logworth False Discovery Rate

* Indicates significance at the 95% confidence level

** Indicates significance at the 99% confidence level

† Ten-fold cross-validation is not supported in JMP with logistic fit; random 90/10 (training/validation) was used instead.

6. Limitations and Future Work

One limitation of this analysis is the relatively small size of the tagged review collection. 3,236 tagged reviews allowed the discovery of significant patterns, but techniques can potentially be refined by tagging a larger review set.

Future work could improve performance by combining the top contributing smoke words from AFINN with manually-curated smoke word lists. In doing so, we may be able to tailor more accurate smoke word dictionaries for dishwasher defect discovery.

Finally, further work is necessary in determining domain-specific sparkle, smoke, and component lists for other industries and product categories, and establishing the extent to which these lists overlap or generalize across industries or product categories.

7. Summary and Conclusions

Our research shows that general sentiment analysis techniques have a modest influence on defect discovery, and domain-specific sparkle and smoke words for the dishwasher product category are critical to efficient automated defect discovery in dishwashers.

In Experiment 1, we found that the AFINN sentiment analysis technique accurately detects defects, but the remaining sentiment analysis techniques that were assessed were outperformed by the unigram, bigram and trigram custom smoke word dictionaries. The smoke-trigram dictionary was the best performing smoke-term dictionary, contained many non-emotive words helpful for defect discovery, and exposed a large number of defects not found by sentiment analysis.

In Experiment 2, we found that logistic regression, neural network, and decision tree classifiers have high performance in defect discovery, when supplied with sentiment, smoke, and sparkle features. In particular, domain-specific "sparkle" terms, which capture user satisfaction and conformance with the product-category-specific design specification or intended usage, are highly influential in all models. Component-specific and outcome-specific features have little impact on dishwasher defect discovery. Finally, neural networks generalize best to holdout folds in the 10-fold cross validation.

Negative online reviews have a demonstrable impact on sales. They can adversely affect brand reputation, and company profitability. Our research extends prior defect discovery research in the automotive, consumer electronics, and toy sectors (Abrahams, et al., 2015; Abrahams, et al., 2013; Abrahams, et al., 2012; Winkler, et al., 2016), introducing new domain-specific dictionaries and classifiers. This research has shown that the dishwasher industry can benefit from tailored "sparkle" and "smoke" term dictionaries which have a powerful impact on bringing to the fore domain-specific dishwasher defects overlooked by general sentiment dictionaries.

APPENDIX A: TAGGING PROTOCOL – DISHWASHERS

This appendix documents the tagging protocol that was distributed to the team of human taggers. The protocol provided the available tag types (attributes), and their definitions, and provided examples for the tagging team of how to tag.

Defect Severity: Tag Type; describes the severity of the defect.

- For defect severity column, tag as follows:
  - *No Defect* – If the consumer is off topic, does not report any issues, or gives only positive feedback.
  - *Safety Defect* – Consumer experiences injury or death by using product.
  - *Performance Defect* – Product does not work as intended by the manufacturer, or as desired by the consumer. If there was an injury or death, use Safety Defect instead!

Injury Timing: Tag type; describes whether or not an injury occurred or could occur, as a result of product usage.

- For Injury Timing column, tag as follows:
  - *No Injury* – If there is no indication of an actual or potential injury as a result of product usage. (If no defect under Defect Severity, above, then Injury Timing must be No Injury)
  - *Minor Injury* – Consumer experiences minor irritations or injury with the product, and no hospitalization was required.
  - *Major Injury* – Consumer experiences severe injury or death with the product, and hospitalization was required.
  - *Potential Injury* – Consumer expresses concern that injury could possibly occur.

Defective Component: Tag type; describes the component of the dishwasher that is labeled as defective by the consumer.

- For Defective Component, column tag as follows:
  - *No Component Defect* – Appliance does not exhibit any specific defective component.
  - *Button Component Defect*^{T} – Dishwasher not functioning properly due to buttons falling off or not responding.
  - *Spray-Arm Component Defect*^{T} – Dishwasher not cleaning dishes correctly due to spray arm not being able to shoot water or reach dishes.
  - *Pump Component Defect*^{T} – Dishwasher not draining properly due to malfunctioned drain pump.
  - *Hose Component Defect*^{T} – Dishwasher not draining properly due to malfunctioned drain hose.
  - *Latch Component Defect*^{T} – Dishwasher does not open or close properly.
  - *Drain/Pump Component Defect*^{S} – Appliance has flooding issues or excess water left in dishwasher due to defective drain or pump component (consumer will usually mention

either or to describe the issue, since both are very closely related/part of the same component).

- *Control Panel Component Defect*[S] – Appliance does not function properly due to a faulty control panel.
- *Rack Component Defect*[S] – Appliance cannot support dishes due to misconfigured or broken upper/lower racks.
- *Motor Component Defect*[S] – Appliance does not function properly due to a defective motor.
- *Soap Dispenser Component Defect*[S] – Appliance's soap dispenser does not open/close properly or does not function as expected by the consumer.
- *Internal Software Component Defect*[S] – Appliance does not function properly due to various software error codes.
- *Internal Hardware Component Defect*[S] – Appliance does not function properly due to faulty circuits, wires, or boards. Also includes flashing or disabled lights.
- *Other Component Defect* – uncommon defects not listed such as faulty fuses, vents, or valves.
- *Multiple Component Defects* – User expresses concerns about multiple components listed above.

Complaint Type: Tag type; describes consumer complaint regarding the dishwashing unit.

- For Complaint Type, column tag as follows:
  - *No Complaint* – User does not complain about any function of the appliance.
  - *Leak Complaint* – User complains of appliance producing excess amount of water inside or outside the unit.
  - *Smell Complaint* – User complains of foul odors emitted from unit.
  - *Noise Complaint* – User complains of too much noise created by unit.
  - *Cleanliness Complaint* – User complains that dishes are not clean or dry.
  - *Time Complaint* – User complains that unit takes too long to clean dishes.
  - *Customer Service Complaint*[S] – User complains that customer service is lacking, or that the company has somehow wronged them.
  - *Cost Complaint*[S] – User complains that the price of the appliance is too high, or that the price of replacement parts and maintenance is costly.
  - *Space Complaint*[S] – User complains that the appliance takes up too much room space (external) or that there is not enough room for dishes inside of the appliance (internal).
  - *Design Complaint*[S] – User complains that the appliance does not meet expectations due to a design flaw (e.g. Most dishes do not fit in between rack spaces, the way the spray arm is positioned causes it to get caught against the rack, control panel placement is vulnerable to electronically harmful environments such as dishwasher steam).
  - *Durability Complaint*[S] – User complains that the appliance has stopped working altogether, has only worked for a few [Day,Month,Year]s, the appliance has inconsistent performances and functions, or has degraded over time.
  - *Other Complaint* – Complaints not listed above.
  - *Multiple Complaints* – User has multiple complaints from the list.

[T]-indicates tag categories used in the Training set only.

[S]-indicates Supplementary tag categories. Tags denoted [S] were added after the initial round of tagging and were not employed in the experiments, but can be leveraged for further managerial insights into the data.

Tags without [T] or [S] indicator were used for all reviews.

Important Notes:

- Use the comment section if:
  - Any of the available tags is not suitable, and/or
  - Further information needs to be provided.
- Be careful when tagging reviews such as "I am very disappointed in the upper rack, as it is not big enough to support my U.S. sized dishes". This is not a "rack defect". This would be considered a "design complaint". Please read the reviews carefully!

Examples of reviews and tags:

Example #1

"Very disappointed with this unit. Have had for two years now. It has new components throughout as they systematically died including main control board, Pump, Soap dispenser, Door latch, Upper and lower racks, Cutlery holder, both spray bars. It does not clean the dishes at all well and I constantly have to baby it along. [Company A] were not remotely interested in talking with us after the warranty expired even though it was the second time the same problems had occurred. I would not recommend this machine to anyone at all."

How to Tag:

- Defect Severity → Performance Defect
- Injury Timing → No Injury
- Defective Component → Multiple Component Defects
- Complaint Type → Cleanliness Complaint

Example #2

"We too have had this dishwasher for 5 years and nothing but trouble: started leaking, repaired, leaking again, motor problems and here's the kicker: we smelt a smoldering and burning smell and then associated it with running the dishwasher; found it was dead. Called the technician who opened the door panel which revealed a complete meltdown of wires and black mush - surely the kind of thing to start a house fire. All [company name] said was, sorry, it`s out of warranty. Do NOT get this dishwasher at any price."

How to Tag:

- Defect Severity → Safety Defect
- Injury Timing → Potential Injury
- Defective Component → Other Component Defect
- Complaint Type → Multiple Complaints

Example #3

"Really gets those dishes, cups, glasses and flatware clean! Excellent performance on pots, pans and casserole dishes too. It has a lot more room inside than I expected, and I only have to run it every other day (family of two). The unit is lightweight enough that this grannie can push it to the sink with ease. It was easy to get out of the shipping container, and the directions to operate it are thorough. What a joy!"

How to Tag:

- Defect Severity → No Defect
- Injury Timing →No Injury
- Defective Component → No Component Defect
- Complaint Type → No Complaint

APPENDIX B: Examples of unique defects found in reviews from the Top 200 Trigrams ranking

Below are excerpts from defects found in the top 200 reviews as ranked using the Trigram-smoke scoring method. The reviews shown here were not in the 200 most-negative reviews in the AFINN sentiment scoring method. Trigrams that matched the trigram-smoke dictionary are shown in **bold**. Notably, a number of the trigrams are non-emotive: they do not explicitly express negative sentiment.

*Two major (expensive) parts **had to be** replaced on October 3, 2008 (**I had to** wait for the parts to arrive; the dealer didn't have the parts in stock) … Several plastic parts came off, and **I had to** snap them back into place. … This **dishwasher does not** do a better job of washing dishes than any of the regular (much less expensive) automatic dishwashers I've owned … I contacted … Customer Service (I made several phone calls) ... First **I was told** that the authorization had been approved and that it had been faxed to the dealer, then **I was told** that the authorization was "still under investigation."*

*I have had this dishwasher for less **than a year** and it doesn't work. **the soap dispenser** does not always open or it opens but the soap is not dispensed … **When I called …** to complain that I **have had the** unit repaired too many times … they said if it is reparable, they **will not replace** it. …*

*I had **to replace the wheels on the** lower basket 3 times, and had it serviced for various things when **the unit was** less than 5. I spent $900.00 when it was new. **The cost of** the parts and service since near $400.00!!*

*The machine does **not dry the** dishes all well … I **end up with** remnants of cereal or mustard or other grossness on my bowls or plates b / c they end up partially covered by other dishes.*

***I should have** kept the old dishwasher! Before the **1 year warranty** ran out, the things that hold **the top rack** to the rollers broke and I called … to get a service tech out there **to fix it** … Then, a few months later, **the control panel** started to act up. … They **came out and** said **the control panel** needed replacing …*

*In the short time I have owned this dishwasher I have had **to replace the** pump and the wheel assemblies for **the top rack**. Last week **the control panel** failed. The combined cost of a new **control panel and** the service call exceeds **the cost of** a new unit.*

*My wife and I purchased this dishwasher a little **over a year** ago and the racks are already starting to rust ... **When I called … I was told** that rust is not covered under warranty.*

***I should have** known. The pump was completely broken on the first dishwasher we purchased and **had to be** replaced immediately. Just **over a year** later, some of the clips **that hold the** tines / racks in place would pop off during the wash cycle … **When I called** to buy new clips, **I was told** the clips to hold the rack in place are not sold separately …*

*I honesty cannot count the times **I had to** take off from week to wait for a service man. The **only good thing** was that customer support did extend the warranty ... Well now it is 6 months out of my warranty and it is not cleaning the dishes. It never did a great job on **the top rack**.*

*… 20 hours **on the phone …** over a $500 dishwasher and all I ever wanted them was to **stand behind their** product.*

*If water gets **on the top** control panel the **dishwasher will not** work for days until it is dry again. Parts keep falling off and it will not run. Don't **waste your money**, it **does not get** better only worse.*

References

Abbott, L. (1956). Quality and competition: an essay in economic theory.

Abrahams, A. S., Fan, W., Wang, G. A., Zhang, Z. J., & Jiao, J. (2015). An integrated text analytic framework for product defect discovery. *Production and Operations Management, 24*, 975-990.

Abrahams, A. S., Jiao, J., Fan, W., Wang, G. A., & Zhang, Z. (2013). What's buzzing in the blizzard of buzz? Automotive component isolation in social media postings. *Decision Support Systems, 55*, 871-882.

Abrahams, A. S., Jiao, J., Wang, G. A., & Fan, W. (2012). Vehicle defect discovery from social media. *Decision Support Systems, 54*, 87-97.

Aggarwal, C. C., & Zhai, C. (2012). *Mining text data*: Springer Science & Business Media.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern recognition, 30*, 1145-1159.

Bradley, M. M., & Lang, P. J. (1999). Affective norms for English words (ANEW): Instruction manual and affective ratings. In: Technical report C-1, the center for research in psychophysiology, University of Florida.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review, 78*, 1-3.

Carvalho, V. R., & Cohen, W. W. (2005). On the collective classification of email speech acts. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 345-352): ACM.

Chakrabarti, S., Dom, B., Agrawal, R., & Raghavan, P. (1997). Using taxonomy, discriminants, and signatures for navigating in text databases. In *VLDB* (Vol. 97, pp. 446-455).

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin, 70*, 213.

Cohen, W. W. (1996). Learning rules that classify e-mail. In *AAAI spring symposium on machine learning in information access* (Vol. 18, pp. 25): California.

CPSC. About CPSC (2015). http://www.cpsc.gov/en/About-CPSC Accessed 04.09.16

CPSC. Recent Recalls (2016). http://www.cpsc.gov/en/Recalls/ Accessed 04.09.16

Crosby, P. B. (1980). *Quality is free: The art of making quality certain*: Signet.

Edward, K., & Stone, P. (1975). Computer recognition of English word senses. In: North Holland Publishers.

Facebook. GE Appliances Company (2016). https://www.facebook.com/geappliances?fref=ts/ Accessed 09.25.15

Fan, W., Gordon, M. D., & Pathak, P. (2005). Effective profiling of consumer information retrieval needs: a unified framework and empirical comparison. *Decision Support Systems, 40*, 213-233.

Fleiss, J. L., Levin, B., & Paik, M. C. (2013). *Statistical methods for rates and proportions*: John Wiley & Sons.

Garvin, D. A., & Quality, W. D. P. (1984). Really Mean? *Sloan management review*, 25.

Godbole, N., Srinivasaiah, M., & Skiena, S. (2007). Large-Scale Sentiment Analysis for News and Blogs. *ICWSM, 7*, 219-222.

Griliches, Z. (1971). Price indexes and quality change.

Hahn, G. J., Hill, W. J., Hoerl, R. W., & Zinkgraf, S. A. (1999). The impact of Six Sigma improvement—a glimpse into the future of statistics. *The American Statistician, 53*, 208-215.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143*, 29-36.

Hernández-Orallo, J., Flach, P., & Ferri, C. (2012). A unified view of performance metrics: Translating threshold choice into expected classification loss. *Journal of Machine Learning Research, 13*, 2813-2869.

Horrigan, J. A. (2008). Online shopping. *Pew Internet & American Life Project Report, 36*.

Johnson, R. M. (1971). Market segmentation: a strategic management tool. *Journal of Marketing Research*, 13-18.

Kamruzzaman, S., Haider, F., & Hasan, A. R. (2010). Text classification using data mining. *arXiv preprint arXiv:1009.4987*.

Kim, S. J., Wang, R. J.-H., Maslowska, E., & Malthouse, E. C. (2016). "Understanding a fury in your words": The effects of posting and viewing electronic negative word-of-mouth on purchase behaviors. *Computers in Human Behavior, 54*, 511-521.

Lancaster, K. (1971). Consumer demand: A new approach.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, 159-174.

Lang, K. (1995). Newsweeder: Learning to filter netnews. In *Proceedings of the 12th international conference on machine learning* (pp. 331-339).

Leffler, K. B. (1982). Ambiguous changes in product quality. *The American Economic Review, 72*, 956-967.

Lewis, D. D., & Knowles, K. A. (1997). Threading electronic mail: A preliminary study. *Information processing & management, 33*, 209-217.

Lipsman, A. (2007). Online consumer-generated reviews have significant impact on offline purchase behavior." comScore. *Inc. Industry Analysis*, 2-28.

McAuley, J., Pandey, R., & Leskovec, J. (2015). Inferring networks of substitutable and complementary products. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794): ACM.

Nagelkerke, N. J. (1991). A note on a general definition of the coefficient of determination. *Biometrika, 78*, 691-692.

Nelson-Peterson, D. L., & Leppa, C. J. (2007). Creating an environment for caring using lean principles of the Virginia Mason Production System. *Journal of nursing administration, 37*, 287-294.

CBS News. 1.7M Maytag Dishwashers Recalled over Fire Risk. (2015). http://www.cbsnews.com/news/17m-maytag-dishwashers-recalled-over-fire-risk/ Accessed 04.09.16

Nielsen. Consumer trust in online, social and mobile advertising grows (2015). http://www.nielsen.com/us/en/insights/news/2012/consumer-trust-in-online-social-and-mobile-advertising-grows.html Accessed 04.09.16

Nielsen, F. Å. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903*.

Ohno, T. (1988). *Toyota production system: beyond large-scale production*: crc Press.

Pak, A., & Paroubek, P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In *LREc* (Vol. 10, pp. 1320-1326).

Riloff, E., & Wiebe, J. (2003). Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 conference on Empirical methods in natural language processing* (pp. 105-112): Association for Computational Linguistics.

Sahami, M., Dumais, S., Heckerman, D., & Horvitz, E. (1998). A Bayesian approach to filtering junk e-mail. In *Learning for Text Categorization: Papers from the 1998 workshop* (Vol. 62, pp. 98-105).

Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., & Demirbas, M. (2010). Short text classification in twitter to improve information filtering. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval* (pp. 841-842): ACM.

Wiebe, J., & Riloff, E. (2005). Creating subjective and objective sentence classifiers from unannotated texts. In *Computational Linguistics and Intelligent Text Processing* (pp. 486-497): Springer.

Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing* (pp. 347-354): Association for Computational Linguistics.

Winkler, M., Abrahams, A. S., Gruss, R., & Ehsani, J. P. (2016). Toy safety surveillance from online reviews. *Decision Support Systems*.

Woodall, W. H., & Montgomery, D. C. (1999). Research issues and ideas in statistical process control. *Journal of Quality Technology, 31*, 376.