IDETC/CIE
International Design Engineering Technical Conferences
& Computers & Information in Engineering Conference
CONFERENCE
August 26-29, 2018
Quebec City Convention Center, Quebec City, Canada

Semantic Classification for Identifying Sustainable Content In Online Product Reviews

CIE 2018 Graduate Research Poster

Nasreddine El Dehaibi
Mechanical Engineering, Stanford University
PhD Student

Erin MacDonald, Assistant Professor, Mechanical Engineering, Stanford University

Motivation

Online product reviews are a viable source for extracting customer preferences but are often unstructured and challenging for designers to gain value from [1]. Multiple studies from literature have shown the use of product reviews for extracting customer preferences [2-5]. This study proposes the use of machine learning techniques to identify sustainable content in product reviews. By extracting customer preferences related to sustainability, this could prove useful for designers in making sustainable products that are successful in online markets.

Methodology

1. Collect Reviews

- **3600 Amazon product reviews**
- **Collected March 2018**

Table 1: Number of product reviews

Product	Number of Reviews
Coffee maker	1258
Lamp	1170
Water filter pitcher	599
Showerhead	232
Paper towels	168
Paper plates	173

Figure 1: Scope of products

2. Manually Label Reviews

- **Qualtrics Survey**

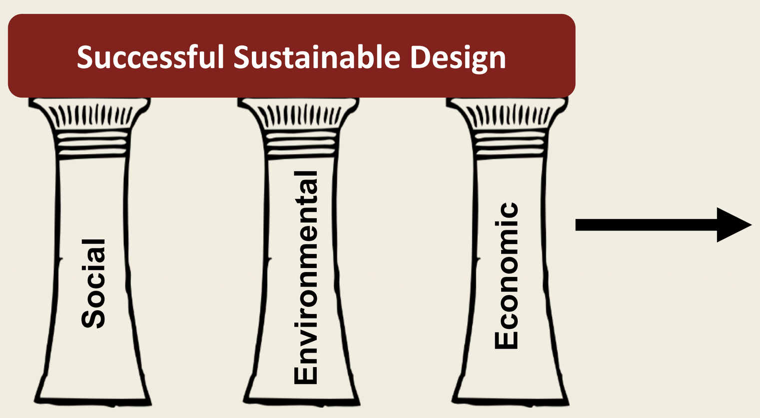

Figure 2: Sustainability Categories

Table 2: Sustainability training

Social Sustainability	Environmental Sustainability	Economic Sustainability
• Health • Education • Safety • Humanitarian efforts	• Durability • Resource consumption • Pollution • End of life disposal	• Affordability for the customer • Business growth • Employment • Profitability

- **MTurk Participants**

Table 3: Survey participant metrics

Number of MTurk participants	200
Number of reviews labeled	3600
Number of reviews approved	3511
Average completion time	20 minutes
Compensation per participant	\$5 + \$2 bonus

- **Labeled Review Example**

Table 4: Example of a labeled review by MTurks

Review	Society	Environment	Economics
Absolutely beautiful design, great output, **low energy use**.	Not relevant	Leans positive	Not relevant

3. Extract Features from Reviews

Table 5: Feature Set

Feature	Options
Cleaning	Lowercase, remove stop words, remove punctuation, stemming
N-grams	Unigrams, bigrams, trigrams
Part-of-speech tagging	Unigrams, bigrams, trigrams
Global Vector (GloVe) Distance	Distance between review and sustainability categories
Amazon Metadata	Product, review rating, review word count
Normalizations	Term frequency inverse document frequency (TF-IDF)

4. Build a Classifier

- Logistic Regression
  - $p(Y=1|X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$
  - $L(\beta_0,\beta_1) = \prod_{i=1}^{n} p(x_i)^{y_i}\left(1-p(x_i)\right)^{1-y_i}$
  - p is probability, L is likelihood, X is feature set, Y is class, β are fitting parameters

5. Evaluate the Classifier

- Split reviews into training and test sets (85%/15%)
- Evaluate using precision $\left(\frac{\text{correct predictions}}{\text{number of predictions}}\right)$, recall $\left(\frac{\text{correct predictions}}{\text{total number of reviews}}\right)$, and F1 (mean score)

Results and Analysis

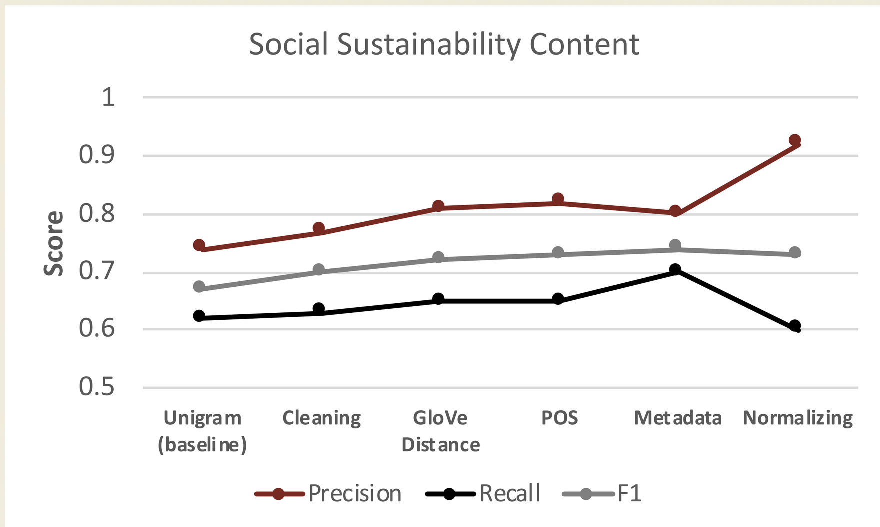

Figure 2: Feature scores for social

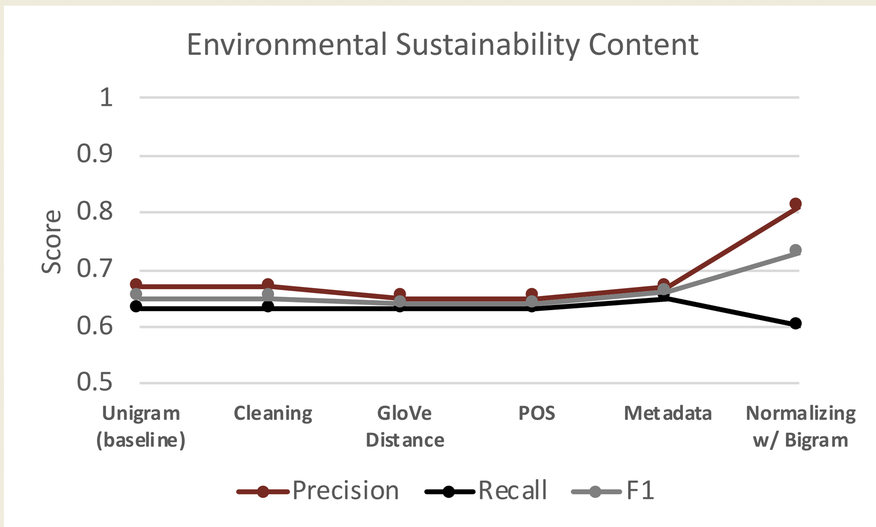

Figure 3: Feature scores for environmental

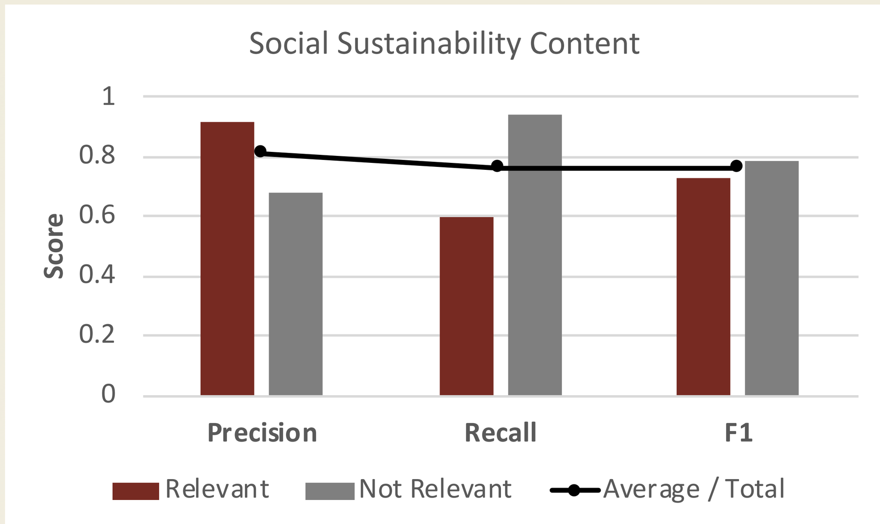

Figure 4: Optimal scores for social

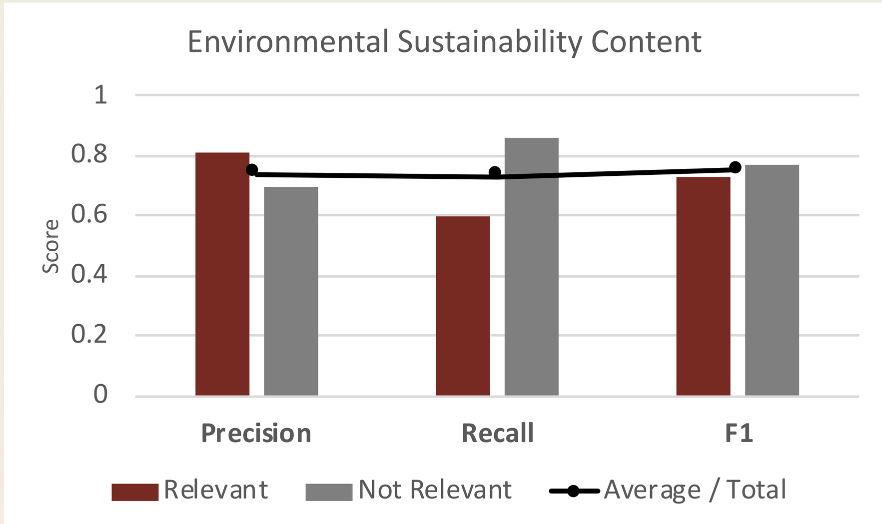

Figure 5: Optimal scores for environmental

Table 6: Percent increase over baseline

	Precision	Recall	F1
Δ	24%	-3%	9%

Table 7: Percent increase over baseline

	Precision	Recall	F1
Δ	21%	-5%	12%

Conclusions

- **Results show potential for modeling social and environmental sustainability in product reviews using machine learning techniques**
- **High precision was achieved for social and environmental sustainability (80-90%) while recall remained insensitive to additional features at 60%**
- **Baseline was outperformed by up to 24% in terms of precision, with review normalizations providing the most significant improvements**

Future Directions

- **Modify the labeling procedure to reduce noise and add product attribute / customer sentiment information**
- **Enhance feature sets to improve recall scores**
- **Identify important product attributes for consumers related to sustainability**

Acknowledgements

We would like to thank Dr. W. Ross Morrow for developing the JavaScript code and server implementation that provides weighted random sampling of product reviews in the Qualtrics survey.

References

[1] Palmer, Stuart 2016, Crowdsourcing customer needs for product design using text analytics, in WCE 2016 : Proceedings of the World Congress on Engineering, International Association of Engineers, Hong Kong, pp. 221-226.
[2] Rai, R., 2012, "Identifying Key Product Attributes and Their Importance Levels From Online Customer Reviews," ASME 2012 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference/ Design Automation, Chicago, IL, August 12 – 15.
[3] Stone, T., and Choi, S.-K., 2013, "Extracting Consumer Preference From User-Generated Content Sources Using Classification," ASME 2013 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference/ Design Automation Conference, Portland, OR, August 4-7.
[4] Tucker, C. S., and Kim, H. M., 2011, "Trend Mining for Product Predictive Design," ASME J. Mech. Des., 133(11), p. 111008.
[5] Tuarob, S., and Tucker, C. S., 2015. "Automated discovery of lead users and latent product features by mining large scale social media networks," ASME Journal of Mechanical Design, 137(7), p. 071402.