Bags *and* Words: Mining Product Features and Customer Sentiment from eBags.com Reviews

Abstract

Product features, along with customer sentiment for those features, are highly desirable to specialty e-commerce retailers so they may provide expertise and value to customers in the areas of product development and curation, shopping search and discovery, personalized recommendations, and effective and efficient marketing. Success in these areas drive retail sales, growth, and customer loyalty in a highly competitive e-commerce marketplace. Written product reviews provided by customers are an authentic and rich source of valuable features and sentiment, but extracting and refining valuable information from reviews is difficult. The goal is to extract product features from raw customer reviews, and attempt to discern customer sentiment for the specific product features. The approach to achieving this goal combines NLP data cleansing and text processing, subjectivity/objectivity partitioning, parts-of-speech (POS) tagging, and sentiment classification using deep learning with minimal supervision to lessen the need for costly and time consuming data labeling.

Introduction

E-commerce is highly competitive with Amazon.com dominating nearly 50% of all online spend in the US in 2018[1]. In order to compete as a specialty retailer, one proven strategy is to demonstrate category expertise by understanding and communicating product attributes and features - and their specific benefits - to prospective customers. Product attributes and features generally are provided by product manufacturers but they are usually just basic, rarely comprehensive, and/or they do not convey valuable benefits that customers may want or discover. Product features may also be further discerned by retailers however this can be time consuming and inconsistent. Further, product features that are problematic or invoke negative sentiment by customers and users are almost never revealed in these activities.

---

1 https://techcrunch.com/2018/07/13/amazons-share-of-the-us-e-commerce-market-is-now-49-or-5-of-all-retail-spend/

It is a common practice for leading e-commerce retailers to solicit and display textual customer reviews that contain product features along with customer sentiment for these features - good or bad. An example of a product feature and customer sentiment from an actual review is:

"I love the bright color inside so you can always find things."

Where the targeted product feature is "bright color inside" and the target sentiment is extremely positive ("love"). The benefit of the feature is also revealed with "you can always find things."

eBags, LLC. is a successful specialty e-retailer featuring the world's top brands in the luggage, bag, and travel accessories categories. eBags has amassed over 3.5 million independent product ratings and reviews. There is potentially very rich product information within these reviews in the form of product features and related sentiment. It would be impossibly time consuming and cost prohibitive to manually read and extract product features and sentiment from textual reviews.

This project proposes successful mining of eBags.com product reviews for product features and benefits, along with an understanding of sentiment of the product features and benefits. The approach to achieving this goal combines NLP data cleansing and text processing, subjectivity/objectivity partitioning, parts-of-speech (POS) tagging, and sentiment classification using deep learning with minimal supervision to lessen the need for costly and time consuming data labeling. If successful, this rich product information can inform product discovery and navigation, product recommender systems, drive targeted marketing campaigns, and even drive product development of valuable features into bags; all helping eBags' customers find their perfect bags and accessories for all of their life journeys and adventures thus enabling eBags.com to successfully compete as a specialty retailer in an Amazon online world.

Related Work and Motivation

Accurately extracting relevant product features from customer reviews while also discerning the customer sentiment toward these product features poses several difficult challenges beyond the common challenges dealing with raw and noisy data. While many reviews contain valuable product features and related sentiment text, they are often buried within large bodies of other

text that is irrelevant to the goal. Once sentences are identified that are more subjective, and thus more likely to have details and sentiment specific to the product, they must be further analyzed to increase the likelihood that they do contain product features and related sentiment, so that information can be further refined and extracted and analyzed. Finally, these refined extractions must be accurately assessed and classified for sentiment. The approach taken to address these challenges relies and builds on work in three key areas; subjectivity/objectivity partitioning, POS tagging, and weakly-supervised deep learning for sentiment classification.

Subjectivity/Objectivity Partitioning

Subjectivity/objectivity partitioning is used to identify sentences that are more subjective and thus more likely to have product features and related sentiment. The technique of (Yeh, 2006) was employed to partition review text into subjective and objective sentences which built on the idea of TexTiling and discourse segmentation (Hearst, 1997). This work proposes that reviews are typically composed of sequences of subjective and objective chunks, and demonstrates a technique of partitioning using a language model based on a subjectivity/objectivity ratio from a corpus of subjective and objective sentences for training that was leveraged by (Pang and Lee, 2004) in their work on subjectivity summarization.

Parts-of-Speech (POS) Tagging

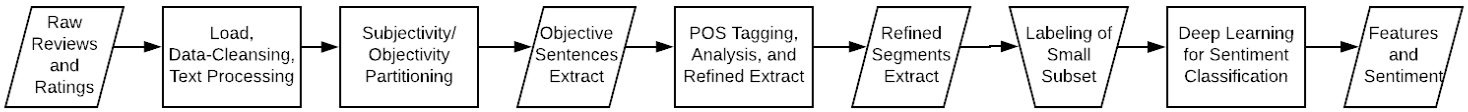
POS tagging and POS pattern analysis is used to further refine and extract text that is more likely to contain product features and related sentiment. This builds on work by (Wang and Ren, 2002) that uses POS patterns to identify interesting feature/attribute and sentiment segments. This is also based on the idea of (Turney, 2002) who also developed POS techniques for extracting phrases for review classification.

Weakly-Supervised Deep Learning for Sentiment Classification

Weakly-supervised deep learning for sentiment classification is used to identify the semantic orientation of the extracted product feature and related sentiment segments with minimal training data. This builds on the ideas and techniques used by (Guan, et. al., 2016).

Methodology

A csv dataset of 369,888 raw reviews was obtained with permission from eBags.com. These reviews are all in the public domain on the eBags.com website and were filtered from the total 3.5 million set by only including active products and also where at least a single rating value was present (in addition to reviews eBags also asks customers to optionally rate their product on a 1-10 scale across the dimensions of durability, price-value, appearance, organization, and overall). The following flowchart illustrates the overall flow and architecture with details for each major step presented below.



Load, Data Cleansing, and Text Processing

Initially, all raw reviews were initially split into sentences using Splitta sentence boundary detection by Dan Gillick[2]. In order to optimize the performance of the Splitta calls, the raw reviews were partitioned into 8 similar sized files so that Splitta could be run in parallel across 8 different Jupyter notebooks (/W266_project/data/SBD_0-Copy*.ipynb). Raw customer reviews can be very noisy sources of natural language and information. Grammatical and spelling errors are common, and html and character codes often litter the data, which make parsing, sentence boundary detection, and identifying basic units of meaning and words difficult. The following techniques were used to clean and process the raw customer reviews; spell check and correction using Peter Norvig's NLP Spell Corrector in Python[3], and html and character code search and replace through regular expressions. Spell correction dictionaries were augmented with an extract of 212,887 word segments representing all manufacturer provided product features from eBags.com, and also augmented with custom dictionary entries to preserve meaningful domain specific terms and jargon (examples; "gripe", and "TSA"). Data cleansing and text processing was done in 1 notebook (/W266_Project/data/Load_DataCleansing_TextProcessing.ipynb). This stage of

[2] https://code.google.com/archive/p/splitta
[3] http://norvig.com/spell-correct.html

the processing flow input 369,888 raw reviews and output 1,069,186 sentences containing 13,752,265 tokens.

Subjectivity/Objectivity Partitioning

The subjective/objective tagged dataset (Pang and Lee, 2004) was retrieved and parsed and was used to train a Naive Bayes classifier with bag-of-words feature engineering. The output of the data cleansing processing was fed into the model for subjectivity prediction and various predictive thresholds were played with to achieve reasonably good output. Initially, the model performed very poorly on many sentences of interest. This was dramatically improved by augmenting the subjective training set with the same set of 212,887 word segments representing all manufacturer provided product features from eBags.com (discussed in the above cleansing step). A subjectivity prediction example of this improvement is shown below:

> "I really like the separate ventilated compartment for shoes and wet swimsuit"
> Subjectivity Prediction: 0.1257

After product feature augmentation:

> "I really like the separate ventilated compartment for shoes and wet swimsuit"
> Subjectivity Prediction: 0.9867

The 132,250 sentences input into this step of the processing flow produced an output of 333,050 sentences that had a subjectivity prediction score above 0.9.

Parts-of-speech tagging

The subjectivity sentence set was run through the nltk tokenize, regex chunking, and POS tagging routines and the patterns (Wang and Ren, 2002) were matched using regex to further refined and split into sentence output of 146,018 sentences that matched and 187,032 that didn't match. Review based trial and error was used to fine tune the patterns so that only sentences with at least 3 tokens would be matched but the same general pattern scheme was used as outlined by Wang and Ren:

Word 1	Word 2	Word 3
JJ	NN/NNS	anything
RB/RBR/RBS	JJ	NOT NN or NNS

JJ	JJ	NOT NN or NNS
NN/NNS	JJ	NOT NN or NNS
RB/RBR/RBS	VB/VBN/VBD/VBG	anything

All of the processing for the Subjectivity/Objectivity Partitioning as well as the Parts-of-Speech Tagging was done in 1 notebook (/W266_Project/product_feature_sentences/Feature_Sentence_Extraction.ipynb).

Weakly-Supervised Deep Learning for Sentiment Classification

The 146,018 "interesting" sentences that remained out of the Parts-of-Speech Tagging step were randomly sampled to produce a small set of 1,039 sentences that were labeled for sentiment of positive or not-positive (neutral/negative/other) to provide input for training and testing on baseline ML and NN models. The premise for such a small labeled set was based on the notion that the Weakly-Supervised Deep (WSD) Learning method used by (Guan, et. al., 2016) could be successfully implemented. Unfortunately this was not the case as various attempts to implement the WSD-CNN model outlined in the paper were unsuccessful. However, an attempt was made to boost sentiment signal with the reviews (in the same spirit of using them for weakly supervised labels), and it appears some success was achieved (see Experimentation and Results section).

Various other models were experimented with for sentiment classification, although with the small training/test sets. A Naive Bayes BoW (NB-BoW) model was used for a baseline and a NLP CNN model, based on a template from Danny Britz[4], was used to test sentiment accuracy. The models were implemented and executed in several Jupyter notebooks with supporting code modules (/W266_Project/deep_learning_sentiment_classification/).

Experimentation and Results

Results can be analyzed over two important objectives:

1. Effectiveness at automated mining and extraction of "interesting" text segments from raw customer reviews, where "interesting" text segments contain product

[4] http://www.wildml.com/2015/12/implementing-a-cnn-for-text-classification-in-tensorflow

feature information and related sentiment.

2. Effectiveness of automated binary sentiment classification on customer review text segments that contain product feature information and related sentiment.

Evaluating the results of the first objective, significant value has been achieved through a technique and model that produces a reduction of 146,018 "interesting" sentences, that contain product feature and sentiment language, from a field of 1,069,186 raw review sentences. This is based on cursory review of the output extract which does indeed look "interesting," however it is worthwhile to attempt to provide more empirical and quantitative validation. To achieve this, a random selection of 1,000 sentences from the initial 1,069,186 sentence corpus was made and labeled as "interesting" or not. This allowed post-facto validation that provided the following estimates produced through comparison and extrapolation with the reduced set of 146,018 sentences:

* Est. post-facto precision = 0.55 (does model let in uninteresting sentences?)

* Est. post-facto recall = 0.15 (does model miss other interesting sentences?)

These results indicate there is still substantial opportunity for refinement and improvement of the technique and model to better achieve this objective.

Evaluating the results of the second objective, strong but questionable accuracy scores were achieved:

Classification Model:	No Ratings Augmentation	Ratings Augmentation
Naive Bayes ML - BoW	0.8421	0.8421
CNN	0.8373	0.8660

These are questionable, and likely overstated, due to two reasons; the small size of the training set, and the imbalanced labeling of the training set. The former was an unfortunate effect of not being able to fully implement the Weakly Supervised approach, which most certainly caused overfitting. The latter, with its 80:20 positive imbalance, likely caused "Accuracy Paradox." Nevertheless, it is interesting to note that when the customer

review scores were added as features by appending them to the end of each sentence, there was a significant improvement in the scores of the CNN model.

Conclusion and Next Steps

This was a valuable NLP project that has and will continue to add real value to the IT, Marketing, and Merchandising teams at eBags.com by making the product feature extraction and sentiment analysis more automated and effective. There are a number of next steps that will be pursued to further improve the techniques, models, and outcome of this project:

- Improve the Subjectivity/Objectivity Partitioning technique and model for better product feature sentence recall and precision. It is expected that fine tuning the parameters and algorithms in the ML model, as well as continuing to add to the Subjectivity training corpus, will help achieve this goal.

- Further analyze and refine the PoS Tag patterns for better precision on extracted features and sentiment - particularly the ability to identify and extract multiple product feature and sentiment segments within sentences (example: "...very easy to find on the carousel with the purple color the spinner wheels are perfect").

- Further assess the accuracy of the sentiment classification models and improve them. It is expected that since only default hyper-parameters were utilized, significant improvement can be achieved with hyper-parameter trial/error/tuning.

- Implement a successful Weakly Supervised Neural Network model that can leverage the weak but important signals of customer ratings to minimize the need for labor intensive manual labeling.

References

[Guan et. al., 2016] Ziyu Guan, Long Chen, Wei Zhao, Yi Zheng, Shulong Tan, and Deng Cai. 2016. "Weakly-Supervised Deep Learning for Customer Review Sentiment Classification." In Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI'16). 3719--3725.

[Hearst, 1992] M.A. Hearst. Direction-based text interpretation as an information access refinement. In P. Jacobs (Ed.), Text-Based Intelligent Systems: Current Research and Practice in Information Extraction and Retrieval. Mahwah, NJ: Lawrence Erlbaum Associates.

[Pang and Lee, 2004] Bo Pang and Lillian Lee, A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts, Proceedings of ACL 2004.

[Turney, 2002] Peter D Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In ACL, pages 417–424, 2002.

[Wang and Ren, 2002] J. Wang and H.Ren, "Feature-based Customer Review Mining," System p.1-9 (2002).

[Yeh, 2006] Eric Yeh. CS224N/Ling237 Final Project Picking the Fresh from the Rotten: Quote and Sentiment Extraction from Rotten Tomatoes Movie Reviews.