Feedback or Research: Separating Pre-purchase from Post-purchase Consumer Reviews

Abstract. Online consumer reviews contain a wealth of information about products and services that, if properly identified and extracted, could be of immense value to businesses. While classification of reviews according to sentiment polarity has been extensively studied in previous work, many more focused types of review analysis remain open problems. In this work, we introduce a novel problem of separating post-purchase from pre-purchase reviews, which can facilitate identification of immediate actionable insights based on the feedback from the customers, who actually purchased and own a product. We address this problem by leveraging state-of-the-art classifiers in conjunction with the features that are based on the dictionaries and part-of-speech (POS) tag patterns. Using the gold standard created from collected online reviews, we experimentally demonstrate that using the features derived from both dictionaries and POS patterns allows all classifiers to achieve higher accuracy for this task than using lexical features alone.

Keywords: Text Classification, Consumer Reviews, E-commerce

1 Introduction

Consumer generated content posted on online review platforms contains a wealth of information, which besides positive and negative judgments about product features and services, often includes specific suggestions for their improvement and root causes for customer dissatisfaction. Such information, if accurately identified, could be of immense value to businesses. Although previous research on consumer review analysis has resulted in accurate and efficient methods for classifying reviews according to the overall sentiment polarity [8], segmenting reviews into aspects and estimating the sentiment score of each aspect [12], as well as summarizing both aspects and sentiments [6] [10] [11], more focused types of review analysis, such as detecting the intent or timing of reviews, are needed to assist companies in making business decisions. One such problem, which we introduce and focus on in the present work, is separating reviews (or review fragments) written by the users after purchasing and actually using a product or a service (which we will further refer to as "post-purchase" reviews) from reviews that are written by the customers who shared their wishes, expectations or results of research before purchasing and using a product (which we will refer to as "pre-purchase" reviews).

Effective separation of these types of review fragments would allow the businesses to better understand what aspects of products and services the customers

are focused on before and after the purchase and tailor their marketing strategies accordingly. It would also allow to measure the extent to which the customer expectations are met by the actual products and services. Furthermore, "post-purchase" reviews, particularly the negative ones, are high-priority reviews, since they provide customer feedback, which needs to be immediately acted upon by manufacturers. Such feedback typically contains reports of malfunctions, as well as poor performance of products that are already on the market. A particularly large number of pre-purchase reviews are created for expensive products that constitute major purchasing decision and require extensive research prior to purchase (e.g. cameras, cars, motorcycles, etc.). There are also many enthusiasts, who often discuss the products they have seen or read about, but do not actually own.

In this work, we evaluate the accuracy of state-of-the-art classification methods in conjunction with the features based on lexical and part-of-speech (POS) patterns for the task of identifying pre-purchase and post-purchase consumer review fragments. Separating these types of review fragments is a challenging task, since it requires distinguishing subtle nuances of language use, identifying implicit clues and making inferences. For example, the past tense of the verb in the phrase "I heard" from the following review fragment "The new Ford Explorer is a great looking car. I heard it has great fuel economy for an SUV" indicates that this positive review has been written by a user, who didn't actually purchase the car. Despite the overall positive sentiment of the fragment, it provides no reliable information to the manufacturer on how the car can be improved. Although the review fragment "so far this is the best car i tested" refers to the past experience, it is a pre-purchase review. On the other hand, while the fragment "If I could, I would have two" refers to the future, it is a post-purchase review. In some cases, the presence of certain keywords gives the clue about the timing of review fragment (e.g. "excellent vehicle, great price and the dealership provides very good service").

In summary, the key contributions of this work are two-fold:

1. We introduce a novel challenging consumer review analysis problem and provide a publicly available gold standard to evaluate the approaches to solve this problem;
2. We experimentally demonstrate that using both dictionary and POS pattern-based features allows classifiers to achieve higher accuracy for this task than using lexical features alone.

2 Related work

Although consumer reviews have been a subject of many studies over the past decade, a common trend of recent research is to move from detecting sentiments and opinions in online reviews towards a broader task of extracting actionable insights from customer feedback. One recent line of work focused just on detecting wishes [9] [5] in reviews or surveys. In particular, Goldberg et al. [5] studied how wishes are expressed in general and proposed a template-based

method for detecting the wishes in product reviews and political discussion posts, while Ramanand et al. [9] proposed a method based on POS patterns to identify suggestions in product reviews. Moghaddam [7] proposed a distant supervision-based method to detect the reports of defects and suggestions for product improvements in online reviews. Therefore, separation of pre-purchase from post-purchase reviews is a novel task that complements these recent studies.

Other non-trivial textual classification problems have been recently discussed in the literature. For example, Bergsma et al. [2] used a combination of lexical and syntactic features to detect whether the author of a scientific article is a native English speaker, male or female, or whether an article was published in a conference or a journal. de Vel et al. [3] used style markers, structural characteristics and gender-preferential language as features for the task of gender and language background detection.

3 Experimental setup

3.1 Gold standard, features and classifiers

To create the gold standard for experiments in this work,[1] we collected the reviews of all major car makes and models released to the market in the past 3 years from MSN Autos[2]. Then we segmented the reviews into individual sentences, removed punctuation except exclamation (!) and question (?) marks (since [1] suggest that retaining them can improve the results of some classification tasks), and annotated the review sentences using Amazon Mechanical Turk. In order to reduce the effect of annotator bias, we created 5 HITs per each label and used the majority voting scheme to determine the final label for each review sentence. In total, the gold standard consists of 3983 review sentences. Table 3 shows the distribution of these sentences over classes. We used unigram bag-of-words lexical feature representation for each review fragment as a baseline, to which we added five binary features based on the dictionaries and four binary features based on the POS tag patterns manually compiled as described in Section 3.2. We used Naive Bayes (NB), Support Vector Machine (SVM) with linear kernel implemented in Weka machine learning toolkit[3], as well as L2-regularized logistic regression (LR) implemented in LIBLINEAR[4][4] as classification methods. All experimental results reported in this work were obtained using 10-fold cross validation and micro-averaged over the folds.

3.2 Dictionaries and POS patterns

Each of the dictionaries contain the terms, which represent a particular concept related to the product (cars, in our case), such as negative emotion, ownership, satisfaction etc. To create the dictionaries, based on discussions and logical

[1] dataset is available at `http://xxxx.xxx/xxx`
[2] `http://www.msn.com/en-us/autos`
[3] `http://www.cs.waikato.ac.nz/ml/weka`
[4] `http://www.csie.ntu.edu.tw/~cjlin/liblinear`

reasoning we came up with a small set of seed terms, such as "buy", "own", "happy", "warranty", that capture the key lexical clues related to the timing of review creation regardless of any particular type of product. Then, we used online thesaurus[5] to find the synonyms of those words and considered each group of words as a dictionary.

Table 1. Dictionaries with associated words and phrases.

Dictionary	**Words**
OWNERSHIP	own, ownership, owned, mine, individual, personal, etc.
PURCHASE	buy, bought, acquisition, purchase, purchased, etc.
SATISFACTION	happy, cheerful, contented, delighted, glad, etc.
USAGE	warranty, guarantee, guaranty, cheap, cheaper, etc.

Using similar procedure, we also came up with a small set of POS tag-based patterns that capture the key syntactic clues related to the timing of review creation with respect to the purchase of a product. For example, the presence of combinations of possessive pronouns and cardinal numbers (pattern "PRP$ CD", e.g. matching the phrases "my first", "his second", etc.), personal pronouns and past tense (pattern "PRP VBD", e.g. matching "I owned") or modal (pattern "PRP MD", e.g. matching "I can", "you will", etc.) verbs, past participles (pattern "VBN", e.g. matching "owned or driven"), as well as adjectives, including comparative and superlative (patterns "JJ", "JJR" and "JJS") indicates that a review is likely to be post-purchase. More examples of dictionary words and POS patterns are provided in Tables 1 and 2.

Table 2. POS patterns with examples.

Pattern type	**Patterns**	**Example**
OWNERSHIP	**PRP$ CD**, PRP VBD, VBZ PRP$, VBD PRP$, etc.	this is **my third** azera from 2008 to 2010 until now a 2012
QUALITY	JJ, JJR, **JJS**	it is definitely the **best** choice for my family
MODALITY	**PRP MD**, IN PRP VBP	buy one **you will** love
EXPERIENCE	VBD, **VBN**	i have **driven** this in the winter and the all wheel drive model

[5] `http://www.thesaurus.com`

Table 3. Distribution of classes in experimental dataset.

Class	# samp.	Fraction
pre-purchase	2122	53.28 %
post-purchase	1861	46.72 %
Total	3983	100 %

Table 4. Performance of different classifiers using only lexical features. The highest value of each performance metric among all classifiers is highlighted in boldface.

Method	Precision	Recall	F1	Accuracy
SVM	**0.734**	0.724	0.717	0.724
LR	0.729	**0.726**	**0.722**	**0.726**
NB	0.703	0.704	0.702	0.704

4 Results and discussion

4.1 Classification of post-purchase vs. pre-purchase reviews using only lexical features

Performance of different classifiers for the task of separating post-purchase from pre-purchase reviews using only lexical features according to the standard performance metrics is shown in Table 4. From the results in Table 4, it follows that LR outperforms SVM in terms of all performance metrics except precision and that both of them outperform Naive Bayes by 2-2.2% on average across all performance metrics.

4.2 Classification of post-purchase vs. pre-purchase reviews using combination of lexical, dictionary and POS pattern features

Results for the second set of experiments, aimed at determining the relative performance of SVM, NB and LR classifiers in conjunction with: 1) combination of lexical and POS pattern-based features 2) combination of lexical and dictionary-based features 3) combination of all three feature types (lexical, dictionary and POS pattern features) are provided in Table 5, from which several conclusions regarding the influence of non-lexical features on performance of different classifiers for this task can be made.

First, we observed that SVM achieved the highest performance among all classifiers in terms of precision (0.752), recall (0.743) and accuracy (0.743), when a combination of lexical, POS and dictionary-based features was used. Second, using POS pattern-based features in addition to lexical ones allowed LR to achieve the highest performance in terms of all metrics and resulted in the highest improvement for NB classifier, while using a combination of lexical, dictionary and POS pattern-based features is more effective for SVM than for both NB and LR. Overall, experimental results presented above indicate that dictionary and POS pattern features allow to improve the performance of all classifiers for the task of separating pre-purchase from post-purchase review fragments relative to using only lexical features.

Table 5. Performance of different classifiers using different combinations of dictionary and POS pattern based features in addition to the lexical ones. The improvement in percentage is relative to using only lexical features by the same classifier. The highest value and largest improvement of each performance metric given a particular feature combination are highlighted in boldface and italic, respectively.

Method	Precision	Recall	F1 score	Accuracy
SVM + POS	**0.733**	0.727	0.722 (+0.70%)	0.727 (+0.41%)
LR + POS	**0.733**	**0.730**	**0.727** (+0.70%)	**0.730** (+0.55%)
NB + POS	0.709	0.710	0.709 (*+1.0%*)	0.710 (*+0.85%*)
SVM + Dictionary	**0.750**	**0.741**	**0.735** (*+2.51%*)	**0.741** (*+2.35%*)
LR + Dictionary	0.740	0.736	0.733 (+1.52%)	0.736 (+1.38%)
NB + Dictionary	0.713	0.714	0.713 (+1.57%)	0.714 (+1.42%)
SVM + POS + Dictionary	**0.752**	**0.743**	**0.738** (*+2.93%*)	**0.743** (*+2.62%*)
LR + POS + Dictionary	0.745	0.741	0.738 (+2.22%)	0.741 (+2.07%)
NB + POS + Dictionary	0.717	0.718	0.717 (+2.14%)	0.718 (+1.99%)

5 Conclusion

In this paper, we introduced a novel problem of separating post-purchase from pre-purchase consumer review fragments, which constitutes an important step towards extracting actionable insights from consumer reviews and found out that combining lexical features with dictionary and POS pattern features improves the performance of all classification models we experimented with for this task.

References

1. L. Barbosa and J. Feng. Robust Sentiment detection on Twitter from biased and noisy data. In *Proceedings of the 23rd COLING*, pages 36–44, 2010.
2. S. Bergsma, M. Post and D. Yarowsky. Stylometric analysis of scientific articles. In *Proceedings of the 2012 NAACL-HLT*, pages 327–337, 2012.
3. O.Y. de Vel, M.W. Corney, A.M. Anderson and G.M. Mohay. Language and gender author cohort analysis of e-mail for computer forensics. In *Proceedings of the Digital Forensics Workshop*, 2002.
4. R.E. Fan, K.W. Chang, C.J. Hsieh, X.R. Wang and C.J. Lin. LIBLINEAR: a library for large linear classification. In *Journal of Machine Learning Research*, 9:1871–1874, 2008.
5. A.B. Goldberg, N. Fillmore, D. Andrzejewski, Z. Xu, B. Gibson and X. Zhu. May all your wishes come true: a study of wishes and how to recognize them. In *Proceedings of the 2009 NAACL-HLT*, pages 263–271, 2009.
6. M. Hu and B. Liu. Mining and summarizing customer reviews. In *Proceedings of the 10th ACM SIGKDD*, pages 168–177, 2004.
7. S. Moghaddam. Beyond sentiment analysis: mining defects and improvements from customer feedback. In *Proceedings of the 37th ECIR*, pages 400–410, 2015.
8. B. Pang and L. Lee. Opinion mining and sentiment analysis. In *Foundations and Trends in Information Retrieval*, 2(1-2), pages 1–135, 2008.
9. J. Ramanand, K. Bhavsar, N. Pedanekar. Wishful thinking: finding suggestions and 'buy' wishes from product reviews. In *Proceedings of the 2010 NAACL-HLT Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 54–61, 2010.
10. I. Titov and R.T. McDonald. A joint model of text and aspect ratings for sentiment summarization. In *Proceedings of the 46th ACL*, pages 308–316, 2008.
11. Z. Yang, A. Kotov, A. Mohan and S. Lu. Parametric and non-parametric user-aware sentiment topic models. In *Proceedings of the 38th ACM SIGIR*, pages 413–422, 2015.
12. J. Yu, Z. J. Zha, M. Wang, T.-S. Chua. Aspect ranking: identifying important product aspects from online consumer reviews. In *Proceedings of the 49th ACL*, pages 1496-1505, 2011.