Predicting the Usefulness of Cosmetic Reviews

Yuri Takashima
Department of Computer Science and Engineering
Toyohashi University of Technology
Toyohashi, JAPAN
yuritaka@kde.cs.tut.ac.jp

Masaki Aono
Department of Computer Science and Engineering
Toyohashi University of Technology
Toyohashi, JAPAN
aono@tut.jp

***Abstract*—Customer reviews, a.k.a. word-of-mouth reviews, have been important resources of information for text mining. They naturally include both positive and negative opinions on the products or services, as well as neutral observations helpful for everyone who is about to purchase the products or about to decide what to do with the product or the service. Among many customer reviews, we focus on cosmetic reviews and propose a machine learning method to predict the usefulness of an arbitrary review. We impose two conditions on the customer reviews. The first condition is that the review should have certain amount of "Likes" given by anonymous users watching the Web site. The second condition is that the time has sufficiently passed since the review has been posted on the internet, in order for us to make sure the votes by the users are at the stage of conversion. We propose a regression model for predicting the usefulness of customer reviews, introducing twenty two features computed in advance by the training data. We also introduce the BoW (Bag-of-Words) model, having more than 8,000 features, as a baseline method, and conduct the comparative experiments. The results demonstrate that the proposed method outperformed the baseline method in terms of RMSE (Root Mean Squared Error) and R-squared. For future work, we expect our proposed features can be applied to predict the usefulness of other customer reviews.**

Keywords—customer review; usefulness; regression; opinion mining

I. INTRODUCTION

Due to the spread and the popularity of customer review Web sites on the Internet, consumers' opinions are amazingly increasing day-by-day. Unlike information sent by companies, diversified opinions and reputations that consumers post on such Web sites are characterized by a large number of "subjective" information. Consumers who utilize information posted on the review Web site are also increasing because they can get honest comments when deciding to purchase, or when using certain products and/or services. For example, the food eating log called "Tabelog" has approximately 1.78 billion PV (Page Views) per month [1], while the cosmetic Web site called "@cosme" has recorded approximately 280 million PV per month [2]. These Web sites prove the high interests in customer reviews among users who keep watching such customer Web sites.

On the other hand, since users can freely post opinions on review sites, the number of reviews on customer Web sites is huge. Users are burdened with browsing all of these reviews and select products or services. Also, the quality of reviews is greatly different. For instance, some users tend to provide very useful information, while other users post reviews that are not concrete. Therefore, if it is possible to extract useful reviews from a huge database, and to provide the reviews with users, they can acquire information on goods and services efficiently, which in turn makes it possible to reduce the burden of users for taking time to view a large amount of reviews.

In this paper, we propose a method to predict the usefulness of the reviews posted on cosmetic Web sites called "@cosme", which is the largest cosmetic Web sites in terms of PV to our knowledge. "@cosme" introduces a system that users can vote for "Likes" if the review is helpful. We consider the number of "Likes" as the indicator of the helpfulness of the review, and attempt to predict the usefulness of an arbitrary customer review. Then, by the structural and semantic analysis of the word-of-mouth, we extract features and construct a regression model from the features as explanatory variables. Using the proposed regression model, we aim to estimate the likelihood of the "Likes" of the reviews.

In the following, we describe an overall system in Section 2, followed by presenting our proposed features in Section 3. In Section 4, experiments and the results are described, and we conclude the paper in Section 5 with some discussions.

II. SYSTEM OVERVIEW

Fig.1 shows the overview of the usefulness judgment system of a word-of-mouth review. The upper part shows the process of preprocessing, including the selection of the word-of-mouth reviews for cosmetic review sites, the extraction of features, and the training of our regression model with the extracted features for judging the usefulness. Specifically, we first collect the reviews from @cosme, and create a word-of-mouth database. Next, by analyzing the structure of the word-of-mouth, we perform syntactic and semantic analysis using morphological analyzers. We then extract features and eventually convert each word-of-mouth into a multi-

978-1-5386-3001-3/17/$31.00 ©2017 IEEE

dimensional vector. For the training, we give the number of "Likes" for each word-of-mouth as the ground truth.

For the testing stage, we give an unknown word-of-mouth review data, and convert it into a vector as with training data. Then, let the system predict the value of "Likes", regarded as the usefulness of the review.

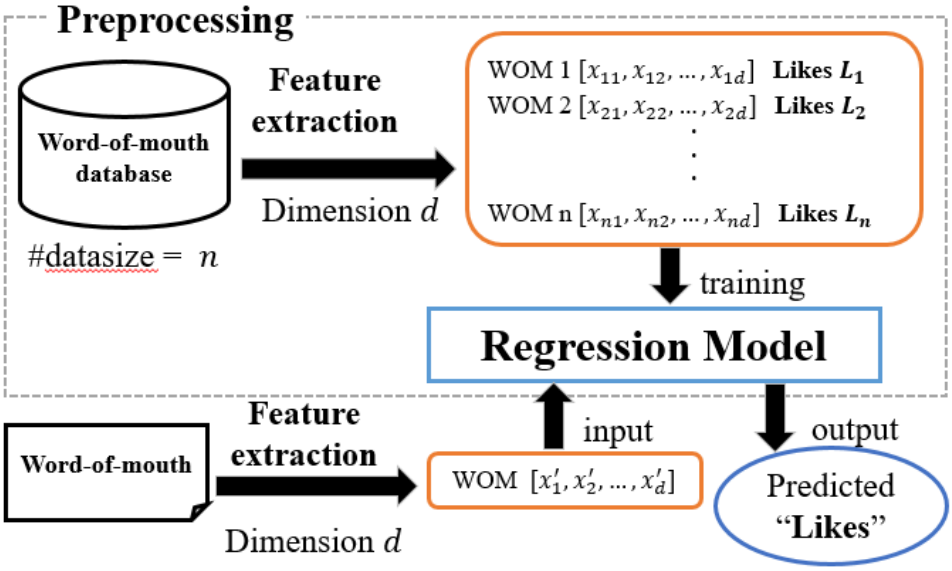

Fig.1 Overall system for predicting the usefulness of cosmetic reviews

III. PROPOSED FEATURES

The proposed features can be classified into three different types: (1) "structural features", kind of meta-data such as the length of reviews and the number of line breaks, (2) "syntactic features", which can be obtained by applying a POS tagger, and (3) "semantic features", which we introduce to reveal the semantic meaning of each word-of-mouth sentence-by-sentence, such as the number of topic changes. In total we propose twenty two dimensional features as summarized in TABLE I.

TABLE I. PROPOSED FEATURES

Types	Features	Dimensions
Structural features	Length of WOM	1
	Number of newlines	1
	Number of ratings	1
	Purchased/Monitored	1
	Actual/Sample	1
	Repeated	1
	Number of images	1
	Number of Exclamations	1
Syntactic features	Number of nouns	1
	Number of verbs	1
	Number of adjectives	1
	Number of adverbs	1
	Number of numbers	1
	Number of technical terms	1
	Number of sentences ending with a noun or noun phrase	1
Semantic features	Polarity	1
	Number of topic changes	1
	Level of topic detail	1
	LDA	4

A. *Structural features*

We propose eight structural features as shown in TABLE I. The first two features as well as the number of exclamations are calculated from the word-of-mouth by simply parsing the text, while the remaining features can be extracted from the meta-data of each word-of-mouth by analyzing the tag patterns. They include the number of ratings, the flag for the items purchased or monitored, the flag for an actual item or a sample, the flag for the presence or the absence of repeat, and the number of images.

B. *Syntactic features*

Syntactic features, which we propose, are based on grammatical observations. Specifically, we propose seven different features. Most of them are obtained by POS (Part-Of-Speech) taggers. We employ the morpheme analysis system MeCab [3] to extract the number of nouns, the number of verbs, the number of adjectives, the number of adverbs, the number of numbers, and the number of sentences ending with a noun or noun phrase.

The number of technical terms in TABLE I is extracted using another POS tagger called Juman++ [4]. Specifically, when performing Juman++, it can mark the out-of-vocabulary (OOV) word as either from an unknown term or from a term defined in Wikipedia, but not in the default dictionary. We observe that the terms marked OOV roughly correspond to technical terms, since the cosmetic technical terms include chemical substances and cosmetic specific terms not often used elsewhere, such as Amazon's customer product reviews. By counting these terms, we propose to add this number to syntactic features.

C. *Semantic features*

We propose four semantic features, including the polarity, the number of topic changes within a review, the level of topic detail, and features extracted from LDA (Latent Dirichlet Allocation) [5].

Regarding the feature for *polarity*, we estimate it by the value obtained from (1).

$$polarity = s(positive) - s(negative) \tag{1}$$

Here, *s*(*positive*) denotes the number of sentences judged *positive*, while *s*(*negative*) denotes the number of negative sentences. The idea behind this equation is simple. If it is in a positive state after taking the difference between the degree of positive and negative, we could conclude that the word-of-mouth is positive, whereas if it is negative, the word-of-mouth is considered negative as a whole. When the degree of positive and negative states cancels out each other, the word-of-mouth may be neutral.

In order to compute *s*(*positive*) and *s*(*negative*), it is necessary to divide each word-of-mouth into a collection of sentences and then attempt to assign the polarity to each sentence. With respect to the process of assigning the polarity to each sentence, we build a sub-system for judging the polarity by SVM (Support Vector Machine) as shown in Fig.2.

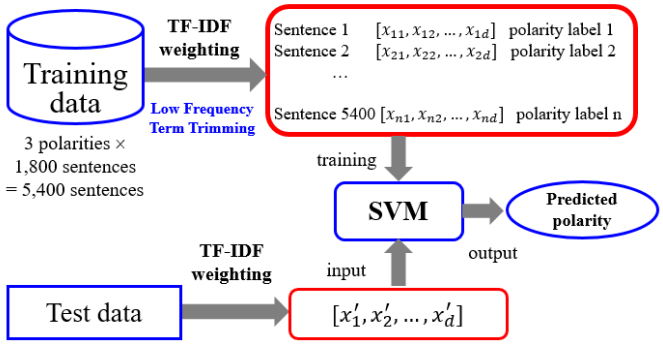

Fig.2 A subsystem for predicting the sentence-based polarity of a customer review

As shown in Fig.2, sentences are classified into three polarities; *positive*, *neutral*, and *negative*. Prior to the training stage, we have made the training corpus, which is a collection of 1,800 sentences with three polarity values, amounting to 5,400 sentences. The judgement for each polarity from the review posted on @cosme, is done manually. Next, we have converted the collected sentences into vector expressions with BoW. For the weighting of BoW, we adopt TF-IDF model. After that, vectorized data is learned by SVM as input data, and a training model is constructed. In the polarity determination unit, a sentence is similarly vector-transformed, and is given as input data to the training model. Then, the prediction label output by the model is regarded as the polarity of the sentence.

The number of topic changes and the level of topic detail, are calculated based on the cosine similarity of adjacent sentences by converting each sentence of the review into a vector expression. The turnover number of topics is a measure showing how many reviews are described in the review. For example, in lotion, there are evaluation items such as moisture and whitening effect, and in general, the more reference items are mentioned, the more useful it is for the reader because it can acquire information of the item. On the other hand, the level of topic detail represents how detailed the description of a topic is. As mentioned above, there are several evaluation items for an item, but the more carefully word-of-mouth is written about the detail of each evaluation item, the more informative it is for the reader on that item.

A method of calculating the number of topic changes and the level of topic detail will be described. First of all, in order to obtain the number of topic changes and the level of topic detail, we divide the word-of-mouth into sentences. After that, the sentence is subjected to morphological analysis with MeCab, and each morpheme is transformed into a vector expression. Therefore, we employ vector expression based on word embeddings generated with Word2Vec [6]. In order to construct a model specialized in cosmetics, we have selected a review of @cosme as the corpus of Word2Vec used here, while the dimension number is set to 100.

The vector representation of a sentence is defined as the sum of each morpheme vector. After converting all the sentences in the word-of-mouth into a vector expression, we calculate cosine similarity between adjacent sentences. It judges that the topic is continuous, if the cosine similarity is less than or equal to the threshold value; otherwise, the topic has changed. The number of topic changes, which is one of the features, is defined as the number of times the cosine similarity becomes equal to or less than the threshold value. On the other hand, the level of topic detail is calculated by (2), where *n* is the number of sentences in the word-of-mouth, while *conv* denotes the number of topic changes above the threshold.

$$detail = \frac{n - conv - 1}{conv + 1} \tag{2}$$

In the feature using LDA, which is one of latent topic estimation methods, we introduce four additional features extracted from the word-of-mouth; the word relevance ratio, the word recall rate, the word precision rate, and the word appearance probability. Using LDA, we try to extract features considering various evaluation viewpoints and topics.

First of all, in order to build a topic model of LDA specialized in cosmetics, we use a word-of-mouth posted on @cosme as a corpus. Next, we obtain a topic group from the model constructed using the corpus. In this model, the number of topics was empirically set to 100. At that time, for each topic group, the LDA model calculates the terms appearing in that topic and their probabilities. Next, using the constructed LDA model, we compute the topic probability as an indicator of each word-of-mouth potentially having the topic within itself. We extract top 5 terms belonging to the topic from the highest estimated probability and make them as feature terms of the word-of-mouth. Finally, we compute the four features; the

number of feature terms in the word-of-mouth, the "term precision" of the word-of-mouth, the "term recall", and the sum of term probabilities. Equations (3) and (4) as shown below define the term precision, and term recall, respectively. Note that c is the number of matches between the terms in the word-of-mouth and the number of feature terms, n is the number of feature terms, and d is the total number of terms in the word-of-mouth.

$$\text{term precision} = \frac{c}{n} \tag{3}$$

$$\text{term recall} = \frac{c}{d} \tag{4}$$

Equation (5) defines the sum of probabilities that match the terms in the word-of-mouth and the feature terms. It should be noted that w_j indicates the term in the word-of-mouth which is matched with the term in feature terms.

$$\sum_{j=1}^{c} \text{probability}\left(w_j\right) \tag{5}$$

IV. Experiment and Evaluation

In this section, we compare the performance of our proposed method for predicting "Likes" of a word-of-mouth against the baseline and discuss the results. We evaluate the proposed method and the baseline with R squared (R^2) and RMSE (Root Mean Squared Error).

A. Dataset

As a result of collecting html documents from @cosme and extracting features from the collected data, we got 88,455 customer reviews. In this experiment, 7,590 customer reviews from January 1st, 2013 to January 1st, 2015 between 5 to 100 "Likes" have been used, where we assume that each review has passed time sufficiently so that "Likes" votes from anonymous Internet used are assumed to be stable.

B. Evaluation scale

As evaluation measures, we choose R squared a.k.a. "contribution rate", and RMSE for the usefulness of a customer review predicted from our proposed judgment system. The contribution rate shows the goodness of fit of our regression model, where the larger the better, while RMSE is an evaluation measure showing how far the predicted value deviates from the correct answer value. The equations (6) and (7) show the details.

$$R^2 = \frac{\left(\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\sum_{i=1}^{n}\left(y_i-\bar{y}\right)\right)^2}{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2} \tag{6}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i-y_i\right)^2} \tag{7}$$

C. Regression model

In this experiment, we have conducted experiments for our word-of-mouth data and verified the results by five-fold cross validation, where approximately 80% of data are used for is training and the remaining 20% are used for testing. SVR (Support Vector Regression) has been used for the regression model. The kernel for SVR is a linear kernel. We have adopted Bag-of-Words model as the feature to be used for the baseline, where each element is weighed with term frequency.

D. Result

The comparative results including the baseline are summarized in TABLE II. As shown in TABLE II, the baseline with Bag-of-Words consists of 8,470 dimensions and our proposed method is 22 dimensions. In spite of lower dimensions, our method outperformes the baseline in terms of R squared and RMSE. We have done experiments with all features, together with Bag-of-Words, as shown in the third row of TABLE II. Finally, we applied feature selection to reduce the dimension from over 8,400 to 250, with the grid search for the parameter tuning, which resulted in the best among all the experiments, shown with underlines and bold faces letters in TABLE II. Experimental results for R squared and RMSE are plotted in Fig.3 and Fig.4, respectively, showing what dimension is most likely to have the biggest (R-squared) and the least (RMSE), respectively.

TABLE II. SUMMARY OF THE EXPERIMENTS

Methods	Dimension	R^2	RMSE
Baseline (BS)	8,470	0.0498	0.1109
Proposed	22	0.1031	0.1083
BS+Proposed	8,492	0.0917	0.1129
BS+Proposed with Feature Selection	250	**<u>0.1520</u>**	**<u>0.1055</u>**

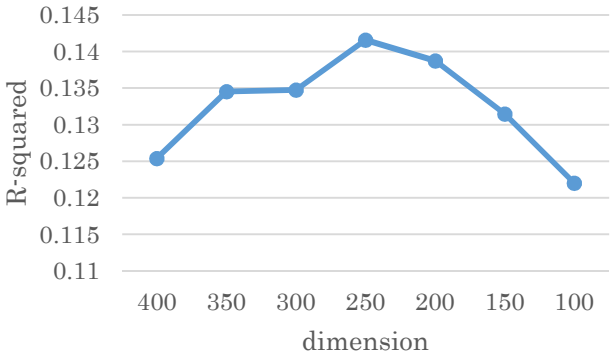

Fig.3 R-squared Graph of the BS+proposed with feature selection, where horizontal axis denotets the dimension. Note that around 250 dimensions, we find a peak.

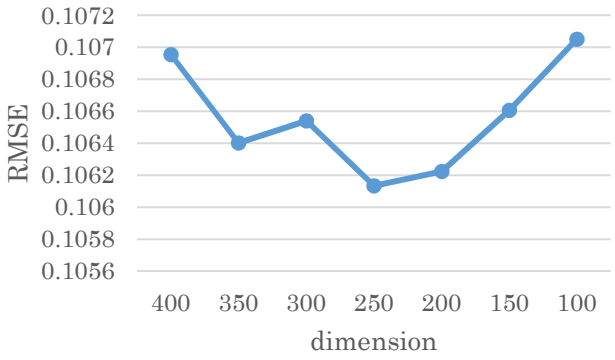

Fig.4 RMSE Graph of the BS+proposed with feature selection, where horizontal axis denotets a minimum error.

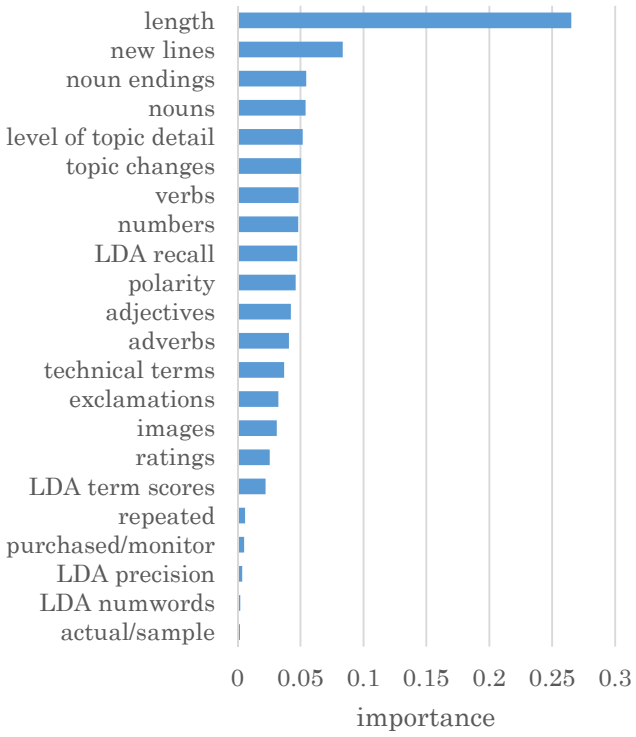

Fig.5 The features sorted by the importance acquired from Random Forest method.

We further conducted experiments using Random Forest regression to investigate what features are effective to predict the usefulness. Apart from several structural features such as the length of word-of-mouth and the number of new lines, some of our semantic features, such as the level of topic detail, LDA term recall, and polarity, turned out to be effective.

V. CONCLUSION AND DISCUSSION

In this paper, we proposed a collection of features to predict the usefulness of a customer review, a.k.a. a word-of-mouth, by taking cosmetic review Web sites as our data, and by assuming that the number of "Likes", given by anonymous Internet users to the customer review, is proportional to the usefulness. The proposed features are structural, syntactic, and semantic features. All together, they consist of twenty-two features. Our especial emphasis has been on semantic features including polarities through our polarity estimation subsystem, the number of topic changes within a customer review, the level of topic detail, and the features extracted from applying LDA, which are term recall, term precision, and the sum of probabilities of terms appearing in the word-of-mouth.

We conducted comparative experiments with a baseline method consisting of 8,470 dimensions of Bag-of-Words features (terms). Our proposed features are 22 dimensions, yet in terms of R squared and RMSE, our proposed method demonstrated better prediction of the usefulness of a customer review. We also conducted experiments using all features (our proposed features + Bag-of-Words), followed by feature selection to reduce to 250 dimensional features, which turned out to be the best.

For future work, we expect our proposed features can be applied to predict the usefulness of other customer reviews. In addition, in this research, we have chosen the customer reviews where the number of "Likes" of each review is more than a certain threshold value. It may be a challenging to predict the number of "Likes" even when the number of "Likes" is below the threshold.

VI. RELATED WORK

In recent years, research of the customer review has been conducted actively. Singh et al. [7] defined the "helpfulness" of reviews posted on Amazon.in and estimated it with machine learning. In all reviews of Amazon, a system has been introduced that allows users to vote whether the review is helpful or not. They defined the helpfulness of the review as the number of votes of "reference" divided by the total number of votes and aimed at estimating the helpfulness by machine learning. As a result, it shows that the semantic feature, related the quality of the review such as readability, polarity, subjectivity and entropy of the review, is important in estimating the helpfulness. Meanwhile, they have confirmed that the structural features such as the length of review, the number of stop words, the number of misspellings did not have much influence.

Hendricks et al. estimated wine characteristics such as color and country of origin from reviews posted on Wine Enthusiast Magazine which is one of wine-specific review sites [8]. The reviews posted on Wine Enthusiast Magazine consist of a small number of words. Therefore, they proposed semantic features using LDA and Word2Vec in addition to Bag-of-Words, and classified wine characteristics by SVM.

The experimental results showed that features combined with LDA and Word2Vec could estimate wine characteristics with higher F-value than features using only Bag-of-Words.

Work using word-of-mouth of @cosme is also widely conducted. Matsunami et al. [9] showed a system for recommending truly helpful word-of-mouth to users. They proposed the system where they first searched similar users in "feeling of use" for the target user to receive a recommendation, and then recommended the word-of-mouth posted by the similar users found in the first step to the target users. For the first step, in order to estimate the similar users in feeling of use, they focused on the rating for each evaluation item of cosmetics, implemented a method to construct a positive / negative expression dictionary for automatically rating word-of-mouth for each evaluation item.

Anbe et al. [10] showed examples of constructing a recommendation system of word-of-mouth according to various purposes of users such as collecting information on cosmetics and searching for new cosmetics. In order to recommend word-of-mouth based on users' interests, they focused on the evaluation viewpoints of the word-of-mouth, and extracted the evaluation viewpoints manually.

ACKNOWLEDGMENT

A part of this research was carried out with the support of the Grant-in-Aid for Scientific Research (B) (issue number 17H01746).

REFERENCES

[1] http://user-help.tabelog.com/advertisement/

[2] http://www.istyle.co.jp/business/uploads/sitedata.pdf

[3] http://taku910.github.io/mecab/

[4] http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN+

[5] David M. Blei et al, "Latent Dirichlet Allocation", *Journal of Machine Learning Research*, 3, pp.993-1022, 2003

[6] https://code.google.com/archive/p/word2vec/

[7] Jyoti P.Singh, Seda Irani, et al., "Predicting the "helpfulness" of online consumer reviews", *Journal of Bus. Res.* 70, pp.346-355, 2017.

[8] Iris Hendricks, Els Lefever, et al., "Very quaffable and great fun: Applying NLP to wine reviews", *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp.306-312, 2016.

[9] Yuki Matsunami, Mayumi Ueda, Shinsuke Nakajima, et al, "Automatic scoring method for item-by-item review using cosmetic item evaluation expression dictionary" (in Japanese), DEIM2016, 2016.

[10] Sayuri Anbe, Ichiro Kobayashi, "Recommendation of cosmetic review based on user attributes" (in Japanese), 22nd Workshop on Natural Language Processing, pp.147-149, 2016.