What Drives Consumer Choices? Mining Aspects and Opinions on Large Scale Review Data using Distributed Representation of Words

Kasturi Bhattacharjee
Dept. of Computer Science
University of California Santa Barbara
Santa Barbara, CA 93106
Email: kbhattacharjee01@gmail.com

Linda Petzold
Dept. of Computer Science
University of California Santa Barbara
Santa Barbara, CA 93106
Email: petzold@cs.ucsb.edu

***Abstract*—With the increasing popularity of online review sites, developing methods to mine and analyze information contained in the vast amounts of noisy user-generated reviews becomes a necessity. In this work, we develop a method to uncover the various aspects of a product or service reviewed by a user, and the opinions associated with them, in an automated fashion. We use the neural network model Word2Vec to build a vector space representation of a large corpus of user-generated, online restaurant reviews, and harness these distributed representations for aspect-based sentiment analysis. User generated text data is intrinsically noisy, with misspellings, informal language, and digressions. Because of the many variations in spelling and expression, the data is also very sparse. Despite these inherent challenges we are able to represent the reviews by key drivers of consumer sentiment, allowing for highly accurate sentiment prediction using a method that is both scalable and human interpretable.**

I. Introduction

With the accessibility and widespread use of the Internet in general, and the rise of social media in particular, user-generated textual content has become pervasive and user opinions are now available freely in the form of reviews on various websites, blogs and comments on social media. Online marketplaces such as Amazon, BestBuy etc. act as a rich source of consumer reviews on the products they sell. Similarly, Yelp and TripAdvisor host millions of reviews on restaurants, businesses, sights, hotels, etc. Consumer feedback is crucial for companies to understand how their products and services are perceived, how they fare in comparison with their competition and help them improve their products and services when the next version is rolled out. Moreover, from the point of view of the consumer, comments and reviews are highly important since learning the opinions of others helps them in their purchase decisions.

Although we have access to large scale user-generated data today, much of the user-generated feedback is in the form of very noisy text from which it is difficult to extract information. Some of the key challenges of working with online text data are: the ambiguity inherent in natural language [1], extreme sparsity [2], [3] and the abundance of noise [4], [5]. Noise may include grammatical errors, misuse of punctuations, spelling errors etc. Individual NLP tasks such as spelling correction, stemming, lemmatization, POS tagging, etc. are often needed to capture signals from such noisy data, which unfortunately, do not scale very well. In addition, specific domain understanding is often required to improve the performance of specific NLP algorithms [6], [7], [8].

To address these challenges, in this paper we present a machine learning approach that utilizes Word2Vec [9], a scalable neural network model that produces a vector space representation of words in order to provide accurate sentiment prediction and human interpretable summary of user-generated online product reviews. The use of this method enables us to extract meaning out of noisy data without having to employ many of the NLP tasks mentioned above.

The motivation of our proposed method comes from the following observation about user reviews. When users review a service or a product, not only do they express their overall opinion on the subject, but they also demonstrate their likes and dislikes over various attributes and functionalities of the service or product in question. For example, when assessing a restaurant, one might like or dislike the food quality, the ambience, the portion sizes and so on. The National Restaurant Association enlists various factors that users consider when choosing a place to eat [10]. Thus, in order to effectively understand why a restaurant is worth eating at or not, it is important to understand these key drivers of sentiment. This leads us into the task of aspect-based sentiment analysis, one of the key frameworks of sentiment analysis today [11], [12], [13], [14]. In the present work, we aim at uncovering the key drivers of sentiment from reviews in an automated fashion, using distributed representations of words, i.e. Word2Vec [9]. We use a publicly available Yelp review dataset [15] for conducting our experiments. We specifically focus on restaurant reviews extracted from this dataset, although the method we propose could easily be extended to reviews on any topic, such as products, vacations, destinations, etc.

Reviews usually contain a numeric rating assigned by the consumer. This rating can be thought of as a mix of positive and negative sentiments that the user feels towards various aspects, details of which may occur in the review text. We use the ratings as labels to train a classifier in order to determine the sentiments associated with the key drivers. Our method performs well across all classifier metrics. Further, using the learned classifier coefficients, we are able to analyze reviews and understand aspects of the topic, i.e. restaurants, that contribute to user satisfaction or dissatisfaction.

Contributions of our work: The main contributions of our work are:

- We develop a method to identify the key aspects of restaurants that are reviewed online and capture the sentiment associated with them. Our method helps in obtaining structure and information from user-generated review data which mostly comprises of noisy, unstructured text.
- Our method provides excellent coverage of the dataset by aggregating contextually similar words, thereby reducing feature space and data sparsity.
- Further, we present in-depth aspect-level analysis of the reviews along with comparative analyses on different kinds of restaurants.
- Although our experiments are conducted on restaurant reviews, the method is generalisable and can be applied to reviews on any service or product.

The rest of the paper is organized as follows: in Section II, we discuss the existing literature on the topic and highlight differences between those works and our proposed method. In Section III, we describe in detail the dataset we use for our experiments and the challenges faced in solving the problem. Section IV provides an outline of the methodology we propose and Section V discusses the details of implementing each step, including the method we compare with. Finally, in Section VI, we illustrate the results obtained using our method and present the Conclusions in Section VII.

II. Related Work

Research in the area of aspect-based sentiment analysis can be broken down into topic modeling based approaches and machine learning based approaches. The topic modeling based methods can be further categorized into two categories - those that separate the task of discovering aspect and sentiment words [11], [13] and those that do not [12], [16]. [12] proposes a flat topic model based on LDA [17], in which a flat mixture of topics is associated with each polarity and all the words with this polarity are generated from this mixture. [13] uses a hybrid model based on Maximum-Entropy and LDA to separately uncover aspect and sentiment words. However, as stated in [18], fully unsupervised models often result in topics that are not always comprehensible by humans, owing to the fact that the objective function used in these topic models does not often correlate well with human judgement.

Outside of the topic modeling framework, Parts-of-Speech (POS) tagging is a widely used method for this problem. The methods proposed in [19], [20] and [21] apply POS tagging to identify nouns and noun phrases, based on the observation that aspects or features are generally nouns [22]. In particular, [19] uses association rules to identify frequent noun phrases, each of which is a possible aspect. In [20], aspects are extracted by computing pair-wise mutual information between noun phrases and a set of meronymy discriminators associated with the product category. Similarly, [21] uses POS tagging along with a language model approach that assumes that product features are mentioned more often in a product review than in generic English.

The above methods are different from ours since none of them use distributed representation of words, and hence, may not capture the contextual similarity between words. However, since POS tagging is a popular method, we use it as a baseline to compare with. We elaborate on the baseline later.

TABLE I

Examples of Reviews from the Dataset. The words in bold indicate noise in the text. Noise includes mis-spellings, case insensitivity, misplaced punctuation marks etc.

Review	**Rating**
My favorite breakfast place. Have good sandwiches also. Stopped again for **Bfast** and had the mixed grill–get the small portion unless you are a real **MAN**! Mixed grill has sausage, (could it be Ricci's?), eggs, onions, and home fries, **soooo goooooooood**! Use Mancini's bread for toast, got the raisin toast - Yum.	5.0
I was first introduced to this place by a friend which ended up being a location we'd frequent when we couldn't decide on where to go, or what to eat. This would be the place we'd hit up for breakfast and on Sundays they have a special brunch menu which offers different items and a buffet style course.	3.0

III. Dataset and Challenges

The dataset we use is a subset of the dataset provided by the Yelp Dataset Challenge [15]. The dataset contains reviews and ratings of businesses as provided by Yelp users, along with meta data consisting of the name and location of the business, the type of the business, etc. To obtain a dataset on a single topic, we extract reviews pertaining to restaurants and thereafter, take a subset of that data. Each review consists of the text of the review, along with the rating that the user provided for that restaurant, which ranges from 1.0 to 5.0. Table I shows some review examples from our dataset.

For the task of sentiment analysis, we use the numeric ratings as a way to label reviews as *positive* or *negative*. On exploring the data, we find that the reviews with ratings 1.0 and 2.0 are mostly negative towards the restaurant under review and those with ratings 4.0 and 5.0 carry positive sentiment. We label reviews with ratings 1.0 and 2.0 as negative, and those with ratings 4.0 and 5.0 as positive. We find that reviews with ratings 3.0 are often ambiguous and hence we omit them as samples. Our dataset consists of 611,696 reviews in all.

Further, we divide the entire dataset using stratified sampling into training (75%) and test (25%) data. This ensures that the rating distribution is retained in both sets. We use the training data for model training purposes as will be subsequently discussed. The test data is used to evaluate our methodology.

Challenges Faced: The following are the main challenges we encounter for this problem:

- The absence of publicly available large scale datasets with annotated aspects and descriptors which makes it a challenge to validate our methods.
- User-generated online data is inherently noisy in nature [4], [5]. Noise includes presence of misspellings, case insensitivity, misplaced punctuation marks, etc. to name a few. Table I shows examples of reviews from our dataset. The noisy words are in bold.
- Users can often be very ambiguous when expressing themselves [1] which makes sentiment analysis difficult as well.
- Data sparsity is another problem, which often arises in analysis of textual data. This is due to the fact that users have different ways of expressing themselves. For instance, some reviews may contain the word *big* to describe a product aspect while others may use synonymous words such as *enormous, huge, gigantic* to express the same idea. In this case, considering each word as a feature leads to high-dimensional, sparse data matrices.

IV. Outline of Methodology

We develop a methodology for automated extraction of key drivers of sentiment from review text, and leverage these drivers in constructing features. These features are subsequently used in a machine learning model for identifying sentiment. In this section, we define and discuss a few key concepts, and present an outline of our proposed methodology.

A. Key Drivers of Sentiment

We aim to identify the aspects that users base their reviews on, as well as the sentiment associated with the aspects. Thus, we propose the identification of the following two groups of words from the reviews:

- **Aspects:** Aspects are the features or attributes of the restaurant under review, such as *food, service, ambience, price*, etc. They form the key elements of the reviews about which users express their likes or dislikes.
- **Descriptors:** Descriptors are words that occur in the neighborhood of Aspects, and either describe the Aspect, or contain underlying sentiment associated with the Aspect. Examples include *tasty, good, disgusting, expensive*, etc.

The following is a review excerpt from our dataset with the Aspects in bold and the Descriptors in italics:

"Let there be no question: Alexions owns the *best* **cheeseburger** in the region and they have now for decades. The **service** is *flawlessly friendly*, the **food** is *amazing*, and the wings? Oh the wings... but it's still about the cheeseburger. The **atmosphere** is *inviting*.... "

As is evident, the review consists of several key aspects of the restaurant the user comments on, such as **food**, **service**, and **atmosphere**. The Descriptor words that accompany these Aspect words carry the sentiment of the user with respect to the corresponding Aspect, e.g., the word *inviting* expresses that the atmosphere of the restaurant was perceived positively by the user.

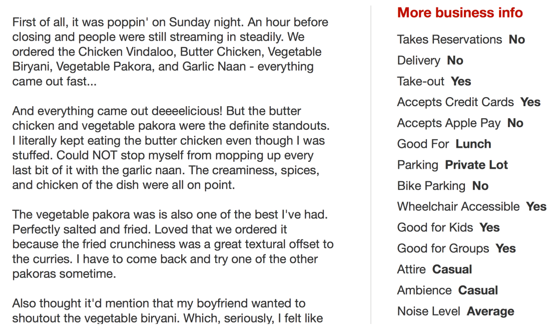

Fig. 1. Snapshot of a restaurant review page on Yelp.com [23]

B. Using Distributed Representation of Words for Automatic Identification of Key Drivers of Sentiment

Evidently, Aspect and Descriptor words are defined by their textual and semantic characteristics, and therefore, we require a method to capture these characteristics from the data. Word2Vec is a deep learning inspired method that generates distributed representations of words based on the contexts in which they occur. The idea behind this concept lies in the Distributional Hypothesis in Linguistics derived from the semantic theory of language use, i.e. words that occur in the same contexts are likely to carry similar meanings [24], [25]. The seminal paper on Word2Vec was written by Mikolov et. al [9], which proposes the Continuous-Bag-of-Words Model and the Skip-Gram model for producing vector representations of words in a corpus. Mikolov et. al use a neural network model with one hidden layer to train word embeddings given the contextual words.

Our motivation for using Word2Vec is to obtain word representations that capture textual, linguistic and contextual characteristics of the words in our data. For instance, words that are synonyms of each other should possess similar representations, while those that are opposite in meaning should not. We would then be able to harness this information in automating the process of identifying the key words. Moreover, this can be of great use for reducing noise which is very prevalent in online text data. Word2Vec has shown promise in a number of previous studies [26], [27] and has been known to capture contextual similarity remarkably well, which is why this was our method of choice.

C. Towards Aspect-Based Sentiment Analysis

1) Building Subgroups Using Contextually Similar Words: In user-generated text, a given concept may often be expressed by different word choices by different users, some of which may even be misspellings. We leverage the Word2Vec model to map all contextually similar words to the same word. Table II illustrates a few such examples. We then define **sub-groups** of Aspects and Descriptors, such that words that are contextually similar are placed within the same sub-group. Table III illustrates some examples of these sub-groups.

TABLE II
INSTANCES OF ASPECT AND DESCRIPTOR SEED WORDS, THEIR MEANINGS AND SOME OF THEIR CONTEXTUALLY CLOSEST WORDS, COMPUTED USING COSINE SIMILARITY. MISSPELLINGS AND INFORMAL LANGUAGE ARE IN BOLD.

Type of Word	Seed Word	Meaning	Contextually Closest Words
Aspect	*food*	Comments on the food and drinks that were served	foods, meals, meal, pizza, cuisine, sushi, burgers, wine, drink
	ambience	Comments on the general environment and vibe of the place.	**ambiance**, atmosphere, environment, vibe, **decore**, setting, layout, interior
	service	Comments on the behavior of the waiter/waitress/bartender/manager and the service received.	**sevice**,services, relations, **svc**, waitstaff
Descriptor	*delicious*	Expressions of the taste of the food served.	**delish**, **delicous**, delectable, **delcious**, **tastey**, tasty
	dirty	Descriptions of the general cleanliness of the place, the food served, etc.	filthy, unclean, smelly, sticky, stained
	professional	Descriptions of the service received from the waiters or the management.	polite, personable, attentive, courteous, hospitable, efficient, respectful

2) Construction of Meta-features: To determine the *sentiment* associated with each Aspect of a restaurant, we propose the construction of **meta-features**. We define these as unordered 2-tuples of the form (a_i, d_j) where a_i represents a word from Aspect sub-group i and d_j represents a word from Descriptor sub-group j, such that the words from d_j occur within a neighborhood m of the aspect word a_i. For example, in the sentence *"I didn't enjoy eating here - the ambience sucks"*, considering $m = 1$, $(ambience, sucks)$ represents a meta-feature that captures the negative sentiment associated with the unpleasant ambience of the restaurant. The goal behind constructing meta-features is two-fold: (1) they help us in capturing the sentiment associated with the Aspects of the reviewed restaurant, and (2) they transform reviews from a large corpus of millions of words to a small set of rich meta features that makes information extraction and analysis easier.

TABLE III
A FEW ASPECT AND DESCRIPTOR SUB-GROUPS OBTAINED USING CONTEXTUALLY SIMILAR WORDS. THESE SUB-GROUPS WERE USED TO BUILD META-FEATURES.

Word Type	Seed Word	Instances of Words in the Subgroup
Aspect	*ambience*	environment, artwork, decour, atmosphere, atmophere, scenery, openness, decoration, atmostphere, decors, decore, vibe, decorations, furnishings
	portion	quantities, portions, quantity, quanity, helping, value, portion, amount, sizing, serving, size
	food	foods, meals, menu, selection, pizza, burgers
Descriptor	*expensive*	pricy, priciest, overpriced, inflated, astronomical, exorbitant, unjustified, outrageous, steep
	clean	sanitary, tidy, spotless, orderly, immaculately, spotlessly , cleaning, cleaned, cleans, neat, squeaky, hygienic
	delicious	tasty, flavorful, delish, delcious, yummers, homemade, onolicious, mouthwatering, addictive,

D. Verification of Proposed Method: Binary Classification

To complete the task of aspect-based sentiment analysis, we must estimate the sentiment-carrying capacity of the meta-features that we determine. In order to do so, we formulate a binary classification problem using logistic regression with l_2 regularization (to prevent over-fitting [28]). Each review acts as a data sample with the class label given by the rating as mentioned in Section III.

In the logistic regression model, $\mathbf{x}_i$ is a data vector of size $k \times 1$ for data sample i, where x_{ij} denotes the frequency of the j^{th} meta-feature in the i^{th} data sample. k is the number of meta-features. y_i is the label of the i^{th} data sample in $\{-1, 1\}$, which is obtained using the numeric ratings as elaborated in Section III.

For the i^{th} sample, the probability that it belongs to the positive class is given by:

$$P(y_i = 1|\mathbf{x}_i, \beta) = \frac{1}{1 + \exp\left(-\beta^T \mathbf{x}_i\right)}, \tag{1}$$

where β is a $k \times 1$ coefficient vector.

V. IMPLEMENTATION DETAILS

In this section, we discuss in detail the implementation of each step of our proposed methodology. As mentioned in Section III, only the training data was used for all the steps of the pipeline up to the Classification step (Steps A to C in this Section). The test data was used in Step D. We use the numerical ratings only for Step D, and use the textual data for the initial steps.

A. Training Word2Vec on Review Data

We use the Python package *gensim* [29] for training Word2Vec, which implements the Skip-Gram model [9]. The input to the model is an ordered sequence of words. The only data pre-processing we perform is to convert the review text into lowercase, to deal with case-insensitivity. Each sentence of a review is tokenized into a sequence of words using Python's NLTK package [30] and fed into the model. There are 3 primary parameters for the model training, namely the word vector dimensions N, the window size w and the minimum frequency count f. N dictates the size of the word embeddings, w determines the size of the neighborhood given a target word, and f represents the minimum number of times a word has to appear in the vocabulary to be a part of model training. We use $N = 150$, $w = 5$ and $f = 20$ in our experiments. After training the model, we now have numerical embeddings of size N for each word in the vocabulary that occurs at least f times.

B. Extracting Key Drivers of Sentiment

To extract Aspects and Descriptors from the reviews, we use the following method:

1) **Determining Aspects and Descriptors**
Aspects**:** We first pick a few seed Aspect words by consulting the Yelp website [23]. Yelp pages containing the reviews of restaurants usually contain a series of features on the right side under "More business info" (Figure 1). These usually contain information on whether the restaurant delivers food, accepts credit cards, has parking, is good for kids, etc. We use these features to create 22 seed words for Aspects, namely *attire, ambience, food, reservations, delivery, payment, cost, portions, taste, service, parking, preparation, celebration, lunch, kids, family, tv, location, clientele, wifi, website, cleanliness*. A few of these Aspect seed words are explained in Table II. The words are chosen such that they span the aspects on which restaurants would be reviewed by users.
Descriptors**:** We explore the neighborhood of Aspect seed words in the training data to obtain Descriptor seed words. For each Aspect seed word in the training data, we extract the co-occurring words (excluding stopwords) from a 5-window neighborhood of the seed word. We then obtain the overall frequency of occurrence of these neighboring words. The 100 most frequently occurring words are manually examined and 21 of them are labeled as Descriptor seed words.
2) **Obtain Sub-groups of Words** For each of the Aspect and Descriptor seed words, we determine their contextually closest words by using cosine similarity on their word embeddings. We use a threshold of 0.5 and select words whose cosine similarity is larger than the threshold. We found the quality of the closest words to drop below that threshold, for most words. Table II contains a few instances of the closest words obtained using Word2Vec. Further, to ensure that each sub-group captures a unique concept and is different from other sub-groups, we unify any pair of sub-groups if the majority of words in either of them are the same. This resulted in merging a couple of Descriptor sub-groups. Thus, we obtain 22 Aspect sub-groups and 20 Descriptor sub-groups.

C. Meta-Feature Construction

To extract meta-features from a data sample, we locate Aspect words (collected in Section V-B) in all sentences of the sample. For every Aspect word, we locate Descriptor words within a neighborhood of 5 words within that sentence. We disregard stopwords during this process.

For example, in the following data sample:
"I'm giving 4 stars mostly because of the beer....large selection & decent prices. The food is pretty good, but nothing to rave about. The menu has a good variety, and everything I've tried has been good. Portions are large.",
(portions, large) would be one such unordered 2-tuple since *portions* is an Aspect and *large* is a Descriptor. Suppose *portions* belongs to Aspect sub-group 1 and *large* belongs to Descriptor sub-group 5. Then, this meta-feature would be indexed $(1, 5)$. If, from a different sentence, we obtain the tuple *(serving, big)*, this meta-feature would also be indexed by $(1, 5)$, since *serving* and *portions* belong to the same Aspect sub-group, and *big* and *large* belong to the same Descriptor sub-group. We have 438 meta-features in all.

D. Binary Classification

Using the meta-features that we construct, we now look for the frequency of occurrence of these meta-features across the training and test datasets, to build our data matrices. There are 438 meta-features, 509,902 training samples and 101,794 testing samples. The label distribution across both matrices is 62.82% positive and 37.18% negative.

VI. Results

In this section, we outline the POS tagging based method we compare with, and present the experimental results obtained using our proposed method. Further, we perform comparative analysis on restaurants, and present those results as well.

A. POS Tagging as a Comparative Baseline

Parts-of-Speech (POS) tagging being a very popular method employed for Aspect-Based Sentiment Analysis [19], [20], [21] we decided to compare our proposed method with a similar pipeline generated using POS tagging. To ensure a fair comparison, we simply replaced the use of Word2Vec in our proposed scheme with that of POS tagging and kept the rest of the pipeline the same. Thus, we still construct meta-features for the comparison, except that we use POS tagging to obtain them. This would enable us to effectively evaluate the necessity of Word2Vec.

Similar to the Word2Vec training approach we adopt, we convert the reviews to lowercase, and tagg our training data using a very popularly used POS tagger, the Stanford POS Tagger [31]. Since Aspects, by definition, are most likely to be nouns, we pulled out the "NN" (nouns) and "NNP" (noun phrases) tagged words from the data. This is similar to the approach taken in [19] for aspect extraction. Further, since Descriptors are most likely to be adjectives, we then look for the presence of "JJ" (adjective) tags within a 5-window neighborhood of nouns. Stopwords are ignored in

TABLE IV
CLASSIFIER METRICS USING l_2-REGULARIZED LOGISTIC REGRESSION

Method	Overall Accuracy (%)	Precision	Recall	Specificity	AUC
Proposed Method	**79.43**	**0.838**	**0.839**	**0.716**	**0.777**
POS Tagging Based Method	67.93	0.792	0.732	0.528	0.63

this process, in the same vein as our proposed method. Each (noun, adjective) pair constitutes a meta-feature, and we collect those that occurred at least 20 times. There are 13,487 meta-features in all. To compare the methods, we then train the same classifier IV-D using these meta-features.

B. Method Validation and Comparison

To quantitatively validate our proposed method, we examine classification results and report the usual classification metrics [32] in Table IV. The metrics reported are accuracy, precision, recall, specificity and AUC. The best results are obtained with the regularization coefficient 0.0001. In order to compare with the POS Tagging based method, we report the same metrics for that method as well. As can be observed, our method performs better than the POS-based method, for all classification metrics. For instance, the overall accuracy obtained using our method is 79.43% whereas that obtained by the POS tagging based method is 67.93%. The main shortcoming of the latter seems to be in capturing negative sentiment, since the specificity achieved is 0.528.

Using the sign and magnitude of each component of the feature weight vector β (1) learned by the classifier, we obtain an explicit sentiment weight for each meta-feature. In Table V, we demonstrate a few meta-features that have the highest positive and negative weights, and were deemed the most discriminative by the classifier. As expected, *(food, delicious)*, *(service, speedy)*, *(price, reasonable)* are all instances of meta features that express positive sentiment, while *(cleanliness, dirty)*, *(food, disgusting)*, *(taste, bland)* convey negative sentiment. Thus, we find positive feature weights to correlate with positive sentiment while negative feature weights correlate with negative sentiment. Making this distinction enables us to do further detailed analyses on users' likes or dislikes about restaurants.

We also compare the meta-features discovered using the POS tagging based method in Table V. It is interesting to observe that amongst the most highly weighted features, the variety of the Aspects captured is very less. Most of the meta-features it discovers are on similar Aspects, e.g. *(wine,wonderful), (coffee, wonderful), (pizza, frozen)* are all references to the food and drinks served in the restaurants.

TABLE V
INSTANCES OF THE MOST POSITIVE AND NEGATIVE META FEATURES UNCOVERED FROM THE FEATURE WEIGHTS DURING CLASSIFIER TRAINING, USING OUR PROPOSED METHOD AS WELL AS USING PARTS-OF-SPEECH TAGS. THE SIGN AND MAGNITUDE OF THE FEATURE WEIGHT VECTOR WAS UTILISED IN OBTAINING THEM.

Method	Sentiment	Meta Features
Proposed Method	*Positive*	(parking, efficient), (attire, classy), (food, delicious), (reservations,speedy), (price, reasonable), (preparation, clean), (delivery, delicious), (service, speedy), (portions, generous), (family, accommodating)
	Negative	(service, disgusting), (delivery, negligent), (food, disgusting), (taste, mediocre), (preparation, disgusting), (cleanliness, disgusting), (wifi, disgusting), (food, dirty), (service, unhelpful), (taste, bland)
POS Tagging	*Positive*	(wine,wonderful), (pizza,wonderful), (bistro, french), (world, top), (coffee, wonderful), (cuisine, great), (tap, great), (pho, phoenix), (diner, welcome)
	Negative	(pizza, frozen), (bread, old), (salad, frozen), (seafood, old), (cheese, frozen), (steak, frozen), (buffet, golden), (chicken, fine), (wings, frozen), (sub, mexican)

Fig. 2. Word Cloud Representing the Most Popularly Used Positive Meta Features. The larger the size of a word, the greater its frequency of occurrence.

In contrast, our method is able to discover a larger variety of review Aspects, even though they may not be correlated, e.g. *(food, delicious), (attire, classy), (cleanliness, disgusting)*. This allows for a wider coverage of consumer sentiments on a variety of subjects. Also, it is to be noted that the meta-features obtained using the POS tagging method are representative tuples of the meta features of our method.

Thus the classifier helps in identification of meaningful, sentiment-carrying meta-features, enabling us to understand consumer sentiment at a more granular level.

Figures 2 and 3 represent word clouds we construct using the occurrence frequency of the positive and negative meta-features. The sizes of the text represent the frequency of occurrence of the meta-features. We use an online tool: WordItOut[33] to build the word clouds. As is evident, meta-features such as *(food, amazing), (service, amazing), (ambience, amazing)* are the most frequently used phrases by people when they express positive sentiment w.r.t restaurants. Similarly, phrases such as *(food, slow), (food, disgusting),*

Fig. 3. Word Cloud Representing the Most Popularly Used Negative Meta Features. The larger the size of a word, the greater its frequency of occurrence.

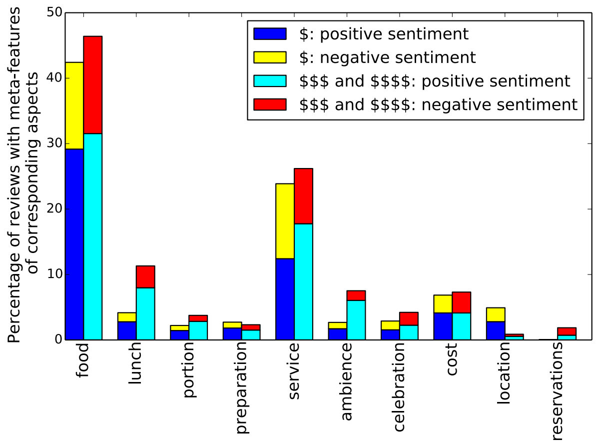

Fig. 4. Comparative Analysis of High-end and Low-end Restaurants. For each Aspect on the x-axis, the y-axis contains the percentage of reviews of the two kinds of restaurants that mention the corresponding meta-features. Differences in color denote different sentiments.

(food, mediocre) reflect the most popular reasons for users to dislike a restaurant.

C. Coverage using Meta-Features

Using the Word2Vec model to construct meta-features has enabled us to capture contextually similar words that may be literally different but semantically similar. This has enabled the coverage of a larger fraction of the data than would have otherwise been possible. For instance, simply looking for the presence of a tuple of Aspect and Descriptor seed words, such as *(food, delicious)* would cover a smaller portion of the entire dataset than looking for the corresponding meta-feature, which captures variations in words, mis-spellings, etc. In Fig 5, we plot the occurrences of a few examples of seed word tokens and those of their corresponding meta-features, to illustrate the increase in coverage we achieve using the meta-features.

D. Comparative Analyses on Restaurants

The sentiment-carrying capacity of the meta-features allows us to perform interesting comparative studies on restaurants, at a granular level. One such study is to compare, on an Aspect-level, what Aspects people review the most and what opinions they express on them when reviewing high-end restaurants vs

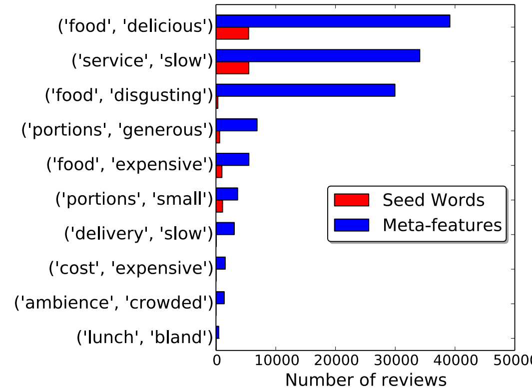

Fig. 5. Difference in Coverage Obtained by using Meta-Features vs Tuples of Seed Words

inexpensive ones. Yelp.com [23] carries information regarding the food prices of restaurants using "$" signs next to the name of the restaurant. "$" implies a cheaper eating place, whereas "$$$" or "$$$$" imply an expensive place. We harness this information when collecting data for the following analysis.

We first extract the 50 most reviewed restaurants present in our dataset, and then determine the low-end and high-end restaurants (as described above) out of those by consulting Yelp.com [23]. Out of these restaurants, we have 11 low-end ones with a total of 33,093 reviews and 9 high-end ones with 23,512 reviews. The rest were "$$" restaurants that we do not consider for the comparison since we were interested in analysing restaurants that lie on two ends of the price spectrum. We then obtain the positive and negative meta-features present in both sets of reviews, Aspect-wise, and plot the ten most popular Aspects and the corresponding aggregated sentiment in Fig 4.

As is expected, ***food*** is by far the single most reviewed Aspect across both kinds of restaurants. ***service*** comes next, and we can observe that the high-end restaurants are slightly more positively viewed in terms of this Aspect. Another interesting and expected observation is that the ***ambience*** is more mentioned in the reviews of the high-end restaurants and is associated with a higher positive sentiment as well. This reflects that clientele of expensive restaurants take into consideration the ambience while such is usually not the case in inexpensive places. Further, cost being the distinguishing factor for the two sets of restaurants here, it is interesting to observe that for the Aspect ***cost***, the number of mentions as well as the proportion of positive and negative sentiment are comparable. The possible explanation for this is that cost is judged on the basis of the food/service received, and not just on the amount of money spent. On closer examination of the reviews, it seems that users are well-aware of the prices of the restaurants they choose, and more often than not are satisfied with the value for money they get. For example, the first two review excerpts are from low-end restaurants and the next two are from expensive ones:

"Was in LV for the wknd was looking for a place to have

dinner that wouldn't cost $$$$ this place was it."

"...In a city (especially the Strip) loaded with overpriced, overcooked and unremarkable food - the Burger Bar is a fabulous find."

"...dinner cost about $230 with tip, which wasn't too bad..."

"...The reviews hating on the cost of bread are out of control. Did you go to Bouchon for a good deal? I hope not. It's expensive. We spent $100 on brunch and honestly thought we got out of there for a steal..."

Thus, a customer of a restaurant, whether expensive or cheaper, is more likely to leave a positive review if she enjoys her overall eating experience, irrespective of the amount of money she spent. As far as other Aspects are concerned, ***reservations*** are discussed w.r.t. the high-end restaurants since these are more likely to require or even offer reservations. Further, ***location*** is discussed more for the inexpensive restaurants since users may not want to go out of their way to eat at these places.

VII. Conclusions

In this paper we demonstrate a method for representing a large corpus of user-generated restaurant reviews by a feature set that captures the *what*, *how*, and *why* of ratings: what aspects customers care most about in a restaurant, how they feel about those aspects, and why. By using contextual embeddings of words we are able to identify and aggregate textual variations with similar meaning, and reduce feature space from 100M tokens to 438 meta-features, achieving strong statistical power while maintaining high coverage of the original corpus. We show that these meta-features have strong predictive power of sentiment, and hence can be used as a way to automatically extract aspect-level feedback from customers automatically and at scale. Our method also enables us to perform comparisons between different kinds of restaurants by analyzing aspect-level sentiment. The method can be extended to other types of reviews as well.

References

[1] A. Franz, "Automatic ambiguity resolution in natural language processing: an empirical approach," *Springer Science & Business Media*, vol. 1171, 1996.

[2] X.-H. Phan, L.-M. Nguyen, and S. Horiguchi, "Learning to classify short and sparse text & web with hidden topics from large-scale data collections," *Proceedings of the 17th International Conference on World Wide Web*, pp. 91–100, 2008.

[3] I. S. Dhillon and D. S. Modha, "Concept decompositions for large sparse text data using clustering," *Machine Learning*, vol. 42, no. 1-2, pp. 143–175, 2001.

[4] L. V. Subramaniam, S. Roy, T. A. Faruquie, and S. Negi, "A survey of types of text noise and techniques to handle noisy text," *Proceedings of The Third Workshop on Analytics for Noisy Unstructured Text Data*, pp. 115–122, 2009.

[5] D. Lopresti, S. Roy, K. Schulz, and L. V. Subramaniam, "Special issue on noisy text analytics," *International Journal on Document Analysis and Recognition*, vol. 12, no. 3, pp. 139–140, 2009.

[6] A. B. Wilcox and G. Hripcsak, "The role of domain knowledge in automating medical text report classification," *Journal of the American Medical Informatics Association*, vol. 10, no. 4, pp. 330–338, 2003.

[7] D. Gruhl, M. Nagarajan, J. Pieper, C. Robson, and A. Sheth, "Context and domain knowledge enhanced entity spotting in informal text," *Lecture Notes in Computer Science*, 2009.

[8] Y. S. Chan and H. T. Ng, "Domain adaptation with active learning for word sense disambiguation," *Annual Meeting-Association for Computational Linguistics*, vol. 45, no. 1, p. 49, 2007.

[9] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[10] 7 factors consumers consider when choosing a restaurant. [Online]. Available: http://www.restaurant.org/News-Research/News/7-factors-consumers-consider-when-choosing-a-resta

[11] Q. Mei, X. Ling, M. Wondra, H. Su, and C. Zhai, "Topic sentiment mixture: modeling facets and opinions in weblogs," *Proceedings of the 16th International Conference on World Wide Web*, pp. 171–180, 2007.

[12] C. Lin and Y. He, "Joint sentiment/topic model for sentiment analysis," *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, pp. 375–384, 2009.

[13] W. X. Zhao, J. Jiang, H. Yan, and X. Li, "Jointly modeling aspects and opinions with a maxent-lda hybrid," *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pp. 56–65, 2010.

[14] L. Zhuang, F. Jing, and X.-Y. Zhu, "Movie review mining and summarization," *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, pp. 43–50, 2006.

[15] Yelp academic dataset challenge. [Online]. Available: https://www.yelp.com/dataset_challenge

[16] S. Brody and N. Elhadad, "An unsupervised aspect-sentiment model for online reviews," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2010, pp. 804–812.

[17] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *The Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[18] J. Chang, S. Gerrish, C. Wang, J. L. Boyd-Graber, and D. M. Blei, "Reading tea leaves: How humans interpret topic models," *Advances in Neural Information Processing Systems*, pp. 288–296, 2009.

[19] M. Hu and B. Liu, "Mining and summarizing customer reviews," *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 168–177, 2004.

[20] A.-M. Popescu and O. Etzioni, "Extracting product features and opinions from reviews," *Natural Language Processing and Text Mining*, pp. 9–28, 2007.

[21] C. Scaffidi, K. Bierhoff, E. Chang, M. Felker, H. Ng, and C. Jin, "Red opal: product-feature scoring from reviews," *Proceedings of the 8th ACM Conference on Electronic Commerce*, pp. 182–191, 2007.

[22] H. Nakagawa and T. Mori, "A simple but powerful automatic term extraction method," *COLING-02 on COMPUTERM 2002: Second International Workshop on Computational Terminology*, vol. 14, pp. 1–7, 2002.

[23] Yelp. [Online]. Available: https://www.yelp.com/

[24] Distributional semantics: Wikipedia. [Online]. Available: https://en.wikipedia.org/wiki/Distributional_semantics

[25] J. R. Firth, "A synopsis of linguistic theory, 1930-1955," *Studies in Linguistic Analysis*, 1957.

[26] T. Mikolov, Q. V. Le, and I. Sutskever, "Exploiting similarities among languages for machine translation," *arXiv preprint arXiv:1309.4168*, 2013.

[27] D. Zhang, H. Xu, Z. Su, and Y. Xu, "Chinese comments sentiment classification based on word2vec and svm perf," *Expert Systems with Applications*, vol. 42, no. 4, pp. 1857–1863, 2015.

[28] A. Y. Ng, "Feature election, l1 vs. l2 regularization, and rotational invariance," *Proceedings of the Twenty-First International Conference on Machine Learning*, p. 78, 2004.

[29] R. Rehurek and P. Sojka, "Software framework for topic modelling with large corpora," *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pp. 45–50, May 2010.

[30] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python*. " O'Reilly Media, Inc.", 2009.

[31] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer, "Feature-rich part-of-speech tagging with a cyclic dependency network," *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, vol. 1, pp. 173–180, 2003.

[32] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

[33] "Worditout." [Online]. Available: http://worditout.com/