Mingxiang Li Yelp Catering Reviews Usefulness Prediction. A Master's Paper for the M.S. in IS degree. April, 2018. 41 pages. Advisor: Jaime Arguello

With the widespread of online businesses, evaluation of customers' feedback is important for the online recommender systems because online reviews have become one of the most important sources of information for modern consumers before purchasing goods or using services. Many recommender systems use user-generated 'usefulness votes' in order to prioritize reviews for users, but there is much room for improvement. In this work, we attempt to predict the the usefulness vote a user will give to the reviews listed in the restaurant category. Using all features, a binary stacked ensemble model achieved a high level of accuracy (0.83). Several feature groups yielded statistically significant improvements while the features related with content don't have great impact to the usefulness. The authors present the results of the study and discuss their significance for research and practice.

Headings:

E-commerce Product Reviews

Text Mining

YELP CATERING REVIEWS USEFULNESS PREDICTION

by
Mingxiang Li

A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Information Science.

Chapel Hill, North Carolina

April 2018

Approved by

____

Jaime Arguello

Content

I Introduction

1.1 Background

Online reviews have received much more attention recently as they have been proven to play an important role to influence the customers' purchase decisions. Reviews are evaluations about various things, ranging from tangible items to intangible items. According to the Cambridge Dictionary, a online review is "a report about a product written by a customer on a commercial website to help people decide if they want to buy it", which is the target of this study.

It has also become more common for costumers to consult related online reviews before purchasing services such as hotels and restaurants. Reading online reviews is the first step in most decision-making processes involving online purchasing (Levi & Mokryn, 2014). Consumers use online reviews to find practical information in daily life, local business operators use them to make profits, and information scientists use them to mine useful information. Online product reviews can be a valuable tool for promoting products, collecting feedback and boosting sales from the marketing perspective (Chu & Roh, 2014; Forman et al., 2008; Hu, Liu, & Zhang, 2008). Characteristics of reviews and reviewers collected through variables such as reviewer identity, reviewer location, information quantity, and semantic factors (Cao, Duan, &

Gan, 2011) may add more insights to the line of research.

While most online review services and retailers rely on peer judgments (e.g., “usefulness votes”) to prioritize reviews for users (Racherla & Friske, 2012), content-based informational cues are also likely to influence the perceived usefulness of a review (Forman, Ghose, & Wiesenfeld, 2008). Online Consumer & Business Review Websites websites, such as Trip Advisior and Yelp, allow all users to evaluate the usefulness of reviews written by others. The online review systems use the number of ‘usefulness votes’ in order to prioritize reviews for users.

In this paper, we focus on the task of predicting the usefulness of Yelp reviews of hotel category by evaluating the effectiveness of a large number of features which include the features that have not been commonly evaluated in prior work.

1.2 Influence of Review

‘‘review helpfulness’’ is particularly important, as it represents the subjective valuation of the review judged by others, and is also the aggregate perceived utility of the information contained in the review (Cao et al., 2011; Baek, Ahn, & Choi, 2012; Li, Huang,Tan, &Wei, 2013). Online reviews have a huge impact on sales. Why people leave reviews after they buy the products or services? Part of the reasons is that the desire to be socially perceived is a powerful motivation for people to leave useful online reviews and can be an important clue to finding useful online reviews (Racherla & Friske, 2012). These reviews are helpful for the costumers and guides them to shop online. Product review usefulness is the subjective evaluation of reviews by their characteristics or ability of providing useful assistance given by peers (Qing, Wenjing, & Qiwei, 2011). Factors that signal the product quality are more important than in offline shopping for decision making (Biswas & Biswas, 2004). When searching for useful information, people often make decisions based on the concepts of quality and authority (Rieh & Belkin, 2000). It is important to know the influence of those factors that signal product quality.

Park and his colleagues claim that (1) the quality of product reviews can improve consumers' purchasing intention, and 2) as the number of reviews increases, the purchasing intention increases as well (Park, Lee, &Han, 2007). The top or easy access reviews are more important because they are correlated with the sales. Previous studies have showed that online consumers are paying particular attention to reviews on the first two pages (Racherla & Friske, 2012). Clemons et al. (2006) found that the strength of the reviews in the top quartile have a positive and significant correlation with sales of microbrewery products.

Even though the latest reviews are helpful, useful reviews will be buried down by chance. As Amazon.com prioritizes the top two most favorable and critical reviews ranked by other consumers, peer ranking of reviews has been regarded as the best method for prioritizing useful reviews (Racherla & Friske, 2012).

However, most of the cases the customers are overwhelmed by the tons of reviews. ‘‘Information overload’’ refers to the variety and quantity of stimuli that exceeds the receiver’s ability to integrate and process them (Jackson & Farzaneh, 2012; Jacoby, 1977; McCormick, 1970). It is difficult for them to select useful reviews and concerns about fake reviews. Information overload impairs comprehension (Hildon, Allwood, & Black, 2012; Lipowski, 1975) and hampers performance (Driver & Mock, 1975; O’Reilly, 1980; Jakoby, Speller, & Kohn, 1974; Schultz, Schreyoegg, & von Reitzenstein, 2013). A review filter is need due to two reasons. First, there are too many reviews and too little time for the costumers. Second not all reviews have “equal” quality. There are “good” reviews from real and unbiased costumers, but there are also “bad” reviews from biased reviewers or auto generators. Most of the time the costumers would prefer to read the reviews in the top or recommended by the system.

One motive of an online review is to influence another person’s behavior in accordance with one’s own preferences, meaning that reviewers may already have their own goals, preferences, and strategic considerations before reviewing a product (Van Rooy, 2003).As a result, product reviewers may not always be maximally rational in their reviews causing variations in their review quality, quantity and relevance. Reviewers may not provide everything they know, as assumed by the cooperative principle (Ganu, Kakodkar, & Marian, 2013; Grice, 1967; Van Rooy, 2003). Some reviewers may even have ulterior motives (Dellarocas, 2003, 2006; Li &

Zhan, 2011; Sotiriadis & van Zyl, 2013). Due to the benefits that extra star ratings can make, many restaurateurs are tempted to leave fake reviews (Anderson & Magruder, 2012). According to the analysis by (Luca and Zervas, 2013), 16% of Yelp users were predicted to be fake users.

1.3 The Object and Assumptions

Our project is an attempt toward identifying useful reviews via predictive modeling based on factors of reviews led by statistical analysis. We build on the prior work and include other features aimed to measure the informativeness of an online review. The model is different with the prior study, stacked ensembles models with Gradient Boosting Machine (GBM) and Random Forest (RF) are used in order to get a better performance.

There are a few hypnoses and assumptions and they are:

H1: the reviews that have more characters, words, or sentences are more helpful because they contain more information regarding the products and user experiences.

H2: rating is a significant predictor of review usefulness.

II Related Work

The quality of the review is measure by the number of useful vote or helpful vote, which is influenced by all the peripheral cues. Korfiatisa and his colleagues defined the quality of the product review as “the number of people who found it helpful out of the total number of people who had read and evaluated the view (Korfiatisa, García-Bariocanalb, & Sánchez-Alonso, 2012).” Qualifications and credibility usually take time to establish. This is the reason online website uses the total review helpfulness votes to determine the quality of reviewers(Albert, Kuanchin, 2015). Using Amazon data, Baek et al.’s (2012) study finds that both peripheral cues, including review rating and reviewer’s credibility, and central cues, such as the content of reviews, influence the helpfulness of reviews.

Research (Keller & Staelin, 1987) has suggested that both quantitative and qualitative factors are relevant to study the quality of information. Quantitative factors of product review refer to the amount, length, volume, and other quantity-related aspects of information. Qualitative factors are more subjective and thus are harder to define. These could refer to the content, writing style, meaning, quality, source, and any other non-quantity aspects of information. Examples of qualitative factors include relevance, accuracy, reliability, timeliness, source credibility, readability, conciseness, sidedness, and others (Alkhattabi, Neagu, & Cullen, 2011; Arazy & Kopak, 2011; Leung, 2001; Wang & Strong, 1996; Yaari et al., 2011).

Many earlier studies and a few recent studies prefer quantified information by using word count, star ratings and etc. Information quantity and quality are interdependent factors. The increase in word count could be an increase in qualitative factors such as relevance and completeness, rather than simply as an increased level of quantity. Prior studies have explored length of the text (Chevalier & Mayzlin, 2006; Gupta & Harris, 2010) and star ratings (Racherla & Friske, 2012; J. Yang, Kim, Amblee, & Jeong, 2012) to predict the usefulness of a review. A study (Mudambi & Schuff, 2010) found a high correlation between the number of words in a review and review helpfulness. It seems that the lengthier of a review, the more likely readers perceive it to be helpful. Another study shows that in most cases, a short review simply does not have the necessary capacity to include all the required elements of a good review (Keller & Staelin, 1987). The finding of the word count can be explained because more words are needed to convey multiple aspects of the detail and thus the review quality increase. The word count is not the only measure for the usefulness: a good review may be filled with details that could make it lengthy, but a lengthy review is not necessarily a good review.

The star rating of the product has been shown to correlate with review helpfulness.

Rating indicates whether reviews of a product or service are positive or negative. When a product is evaluated positively, the product has high product rating. A review that rates a product five stars would logically contain more positive information than the review that rates only one-star (Mudambi & Schuff, 2010; Poston & Speier, 2005). Some readers depend more on low-rating reviews because they feel that they are more diagnostic and thus more useful (Ahluwalia, Burnkrant, & Unnava, 2000), and positive reviews as less helpful because of their weaker perceived depth (Hao, Li, & Zou, 2009). There is also evidence that product ratings are positively associated with review helpfulness (Mudambi & Schuff, 2010). These studies show, there is a direct relationship between product ratings and sales.

The value of a review assessed by a user increases when it provides more information (Daft & Lengel, 1986). Pang et. al. focus on the thumbs up and down to classify, analyze, and rank the quality of the reviews (Pang, Lee, & Vaithyanathan, 2002). Li et al. (2013) conducted a study and found that the content-based review features have a direct impact on product review helpfulness. Consumers perceive customer-written product reviews as more helpful than those written by experts. A customer-written product review with a low level of content abstractness yields the highest perceived review helpfulness.

There are much more qualitative information research in recent studies. Attributes of the review are more likely to influence its usefulness. The content-based features studies are list as follow: writing style and timelines.(Liu, Huang, An, & Yu, Modeling and Predicting the Helpfulness of Online Reviews, 2008), message sidedness and extremeness (Cheung, Luo, Sia, & Chen, 2009; Schlosser, 2005), vividness and strength of the message (Sweeney, Soutar, & Mazzarol, 2008), sentiment (Levi & Mokryn, 2014; Sweeney, Soutar, & Mazzarol, 2008), amount of information (Chevalier & Mayzlin, 2006), and organization/structure of information presentation (Rieh, 2002).

The quality of information is crucial in online reviews, as high quality information provides reliable, current and concise information (Arazy & Kopak, 2011; Yaari, Baruchson-Arbib, & Bar-Ilan, 2011). In the online review context, quality of information relates to the qualifications and credibility of reviewers (Li & Zhan, 2011;Sotiriadis & van Zyl, 2013). High-quality information might be characterized as accurate, reliable, current, concise, fair, easy to understand, organized, and many other things (Alkhattabi et al., 2011; Arazy & Kopak, 2011; Yaari et al., 2011).

Sentiment information also help in filtering useful review. The framing of a message may be neutral, positive, or negative toward a product, and may affect the perceived value of the message (Grewal, Gotlieb, & Marmorstein, 1994). Cao et al. (2011), employing data from CNET Download, state that the semantic characteristics are more influential than other characteristics in affecting how many helpfulness votes reviews receive. Reviews with extreme opinions receive more helpfulness votes than those with mixed or neutral opinions. Mudambi and Schuff (2010) conclude that review extremity, review depth, and product type affect the perceived helpfulness of the review. Review depth has a positive effect on the helpfulness of the review, but the product type (search or experience) moderates the effect of review depth on the helpfulness of the review. The relationship between framing and consumer attitude have been conflicting: Some studies suggest that negatively framed information is more credible because it is unlikely to be contributed by the product's sellers or manufacturers (Kanouse, 1984). Some people may prefer messages that present both the negative and positive aspects of a product (Hastak & Park, 1990), and so may view online reviews that contain both the pros and cons of a product to be more objective and thus more believable. Two sided arguments were also more persuasive than one-sided positive arguments when the initial attitude of the consumer was neutral or negative (Crowley & Hoyer, 1994). Some studies have also shown that reviews that carry both positive and negative opinions receive more helpfulness votes than those with neutral opinions (Cao et al., 2011).

Different models are used in different studies. Pan and Zhang used a mixed effect logistic model with random intercepts due to the characters that logistics model is appropriate for the binominal distribution of data ("Was this review helpful to you? Yes or No?")(Pan & Zhang, 2011). Logistic regression is intrinsically simple, it has low variance and is more robust: the independent variables don't have to be normally distributed, or have equal variance in each group, so is less prone to over-fitting. Kim describes a system that can rank Amazon product reviews based on helpfulness using SVM regression, and the paper also presents an in-depth analysis of the importance of the structural, lexical, syntactic, semantic, and meta-data features to review helpfulness.(Kim, Pantel, Chklovski, & Pennacchiotti, 2006) The SVM regression tool “SVMlight” is used and the trained SVM model automatically return the score and rankings based on the list of features selected. There are three main advantages for the SVM: First it has a regularisation parameter, which avoids over-fitting. Second it uses the kernel trick, a mapping function. Third, SVM is defined by a convex optimisation problem for which there are efficient methods.

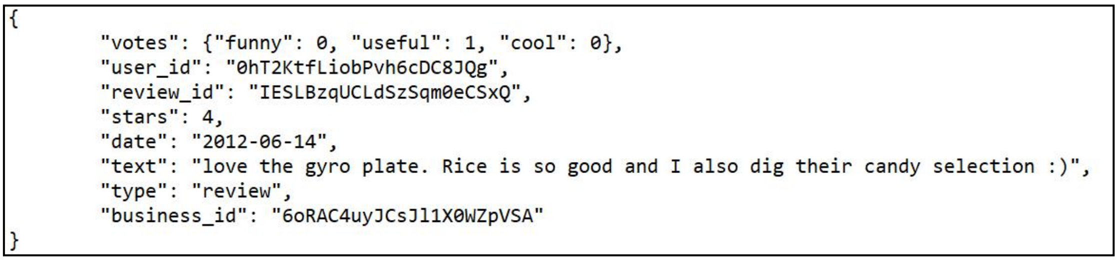
In statistics and machine learning, ensemble methods use multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms alone.(Opitz, Maclin, 1999). Ensemble methods have been called the most influential development in Data Mining and Machine Learning in the past decade. They combine multiple models into one usually more accurate than the best of its components. Ensembles can provide a critical boost to industrial challenges where predictive accuracy is more vital than model interpretability. Ensembles are useful with all modeling algorithms(Giovanni, John, 2010). Empirically, ensembles tend to yield better results when there is a significant diversity among the models (Kuncheva, Whitaker, 2003).

Most of the studies that focus on studying usefulness (or helpfulness) in online reviews used Amazon reviews (Ghose & Ipeirotis, 2011; Kim, Pantel, Chklovski, & Pennacchiotti, 2006), Yelp Reviews (Levi & Mokryn, 2014; López & Farzan, 2014; Pentina, Bailey, & Zhang, 2015; Racherla & Friske, 2012) or other sources (Lu, Tsaparas, Ntoulas, & Polanyi, 2010).

III. Data

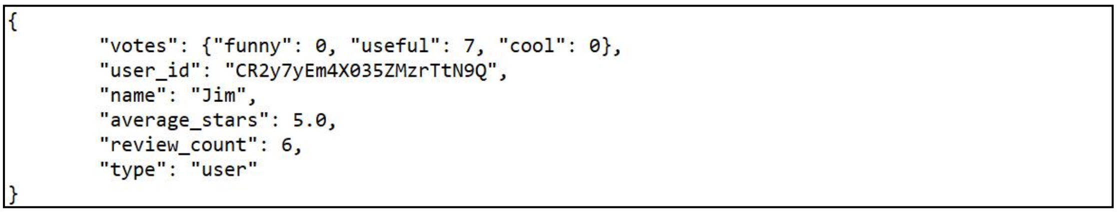
We utilized the dataset provided as part of Yelp Recruiting Competition. Yelp is the largest business listing site for service businesses. Due to the large data size, Yelp is considered as a representative online review service when considering service review part only (Racherla & Friske, 2012). The dataset contains online reviews of Arizona State local businesses written between 2013-01-19 and 2013-03-12. There are 229,907 reviews writen by 43,873 users for 11,537 local businesses.

Data for review, reviewer, and business were provided as JSON objects in separate files. The following are the snapshot of the review, reviewer, and business JSON files.



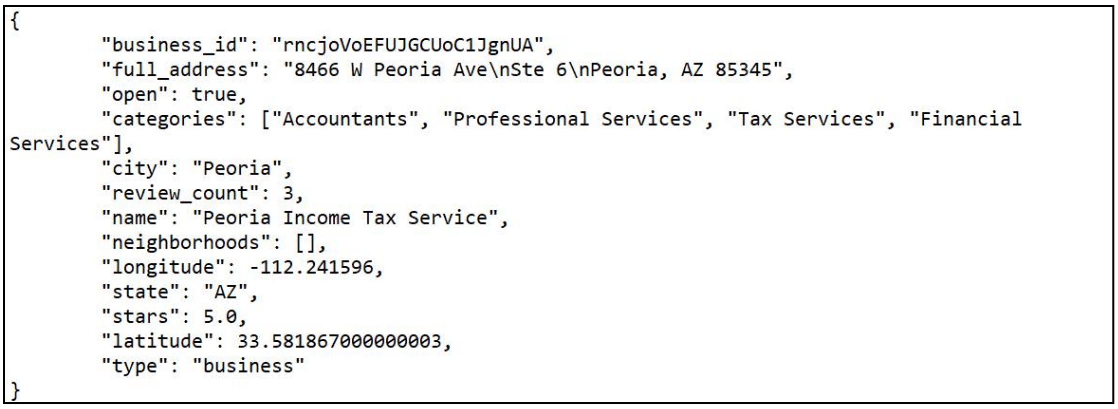
```
{
        "votes": {"funny": 0, "useful": 1, "cool": 0},
        "user_id": "0hT2KtfLiobPvh6cDC8JQg",
        "review_id": "IESLBzqUCLdSzSqm0eCSxQ",
        "stars": 4,
        "date": "2012-06-14",
        "text": "love the gyro plate. Rice is so good and I also dig their candy selection :)",
        "type": "review",
        "business_id": "6oRAC4uyJCsJl1X0WZpVSA"
}
```

Figure 1: Review JSON files



```
{
        "votes": {"funny": 0, "useful": 7, "cool": 0},
        "user_id": "CR2y7yEm4X035ZMzrTtN9Q",
        "name": "Jim",
        "average_stars": 5.0,
        "review_count": 6,
        "type": "user"
}
```

Figure 2: Reviewer JSON files

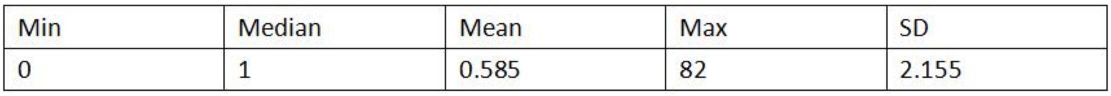


```
{
        "business_id": "rncjoVoEFUJGCUoC1JgnUA",
        "full_address": "8466 W Peoria Ave\nSte 6\nPeoria, AZ 85345",
        "open": true,
        "categories": ["Accountants", "Professional Services", "Tax Services", "Financial
Services"],
        "city": "Peoria",
        "review_count": 3,
        "name": "Peoria Income Tax Service",
        "neighborhoods": [],
        "longitude": -112.241596,
        "state": "AZ",
        "stars": 5.0,
        "latitude": 33.581867000000003,
        "type": "business"
}
```

Figure 3: Business JSON files

These JSON files include attributes that can be directly used as features and also has attributes that can be converted through processing.

The number of "useful votes" associated with each Yelp review is used as the standard for judging whether a review is useful. In this work, we decided to quantify the usefulness prediction task as binary classification. It was necessary to determine how to binarize the data into "useful" and "not_useful" labels using the number of usefulness votes. The distribution of the number of usefulness votes is list as follow:



Min	Median	Mean	Max	SD
0	1	0.585	82	2.155

Table 1: Distribution of the number of usefulness votes

The average is 0.585 and the median is 1. For the experiments, in order to make the data distribution balanced(close to 50%vs50%), the threshold ζ was set to 2. There are 37% usefulness reviews and 63% unuseful reviews according to this threshold. Reviews with two or more "usefulness votes" were considered useful and reviews with less than 2 “usefulness votes” (0 or 1) were considered not_useful.

One important factor that might influence the number of “usefulness votes” is the review’s exposure time which is related with the date attribute. Older reviews have more chance to accumulate “usefulness votes” than the newer ones. In this experiment, we decided to sample reviews written during the 2 years: 2011 to 2013 in order to remove the influence of potential confounding factor.

This project is only focusing on the catering industry, 92276 reviews are used in the final dataset.

IV. Features and Models

Features utilized in this study are grouped into 7 categories: basic features, sentiment, date, text length, capital letter, punctuation count and rate features. All the features are extracted from review, reviewer and business files. We utilized 60 features in the experiment.

4.1 Features

Basic features(5)

Basic features include rate, business id, word count, number of the capital letters and punctuation count. Rate and business id are the raw data contain in the review files. Rate is the score the reviewer gives to a business and is an integer value from 0 to 5 indicating how many “starts” in average product reviewers give to a product. Rate is a direct and simple evaluation for the products or services received in a business. Rate is like an abstract of the review. It’s an important factor in the usefulness judgement because people might agree with someone’s post and click the useful vote without reading all the text but they tend to read the rate first and not miss it. We are curious to see if there is any relationship between score ratings and review usefulness. The rate might have a positive effect before the user read the reviews. Such positive perception might affect the audience to endorse those positive reviews about the successful products and undermine the negative product reviews' usefulness. Business id means which restaurant the user comment about. It is use to differentiate the influence of the business.

The features with little processing is extracted from the text field of the review file. There are 3 features: word count, number of the capital letters and punctuation count. As we hypothesize, longer reviews may contain more information about the products, which the reviewers might find useful. Obviously, a long review tends to have more words, capital letters, punctuation and sentences. Thus, we invite these factors to the party.

The word count shows how many words are contain in the reviews. It is a content based feature and is often use as a measurement for the informativeness. More word count might contain more information regarding the products and user experiences. The number of capital shows how many capital letters are contain in the reviews and it might have some relationship with important words. Capitalization might be used for proper nouns, specific regions, specific business, the first word of a sentence or something else. Punctuation count shows how many punctuation is include in the review and punctuation is used to create sense, clarity and stress in sentences.

Punctuation might relate with the informativeness. More punctuation a review has, more likely that it contains more information. Punctuation might also be able to represent the polarity of the reviewers' sentiments when they write the reviews. We expect the reviews with more extreme emotions to be deemed non-helpful because of lack of subjectivity.

Punctuation and capital are part of the readability features, they can both improve or decrease the quality of writing style, review structure and informativeness of the review.

Sentiment Features(5)

Sentimental features are generated from the text field of review file. The emotion detection method is bag-of-words. The tool we use is VADER Lexicon. VADER stands for Valence Aware Dictionary and Sentiment Reasoner. It is a lexicon with a rule-based sentiment analysis framework that was specially built for analyzing sentiment from social media resources (Gilbert, C. H. E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text). There are 5 measurements of VADER Lexicon: positive, negative, neutral, compound score, and final sentiment. The positive score, negative score and neutral score are ratios for proportions of text that fall in each category and they sum up to be 1. The threshold values of final sentiment are:

1. positive sentiment: compound score >= 0.05

2. neutral sentiment: (compound score > -0.05) and (compound score < 0.05)

3. negative sentiment: compound score <= -0.05

Date features(10)

The date features can be divided into 2 branches: the date of the business's all reviews and the date of the user's all reviews. We calculate the average, median, minimum, maximum and standard deviation value of the date features in order to see whether the distribution of data influence the result. There are 5 date features for a business and 5 date feature for a user.

To calculate the average date, first we set a standard such as 2010/01/01. Then we calculate the difference day between the standard and the date of the review. We add all the differences and divide the count to get the average. We finally get the average date by adding the average number to the standard. The median date is calculate by the similar method, which shows the median date of all the reviews a business receive after they register in Yelp. Minimum date means the date of the first review a business receive and the maximum date means the date of the latest review. Standard deviation is used to quantify the amount of variation of all the date data of a specify user or business.

The average, median, minimum, maximum date feature is in

XXXX(year)-XX(month)-XX(day) form and the results are rounding. The standard deviation of the date is represent by number.

The reason why the date data is used is that the activity of the user such as posting reviews and the activity of business such as receive reviews can be shown in the above five metrics. The date data of an account include the activities information and the data distribution is useful in judging the review post/receive frequency.

Text length features(10)

The text length features can be divided into 2 branches: the text length of the business's all reviews and the text length of the user's all reviews. The text length features here are correlated with the user and business, so it's different with the text length feature in the basic feature section. We calculate the average, median, minimum, maximum and standard deviation value of the text length features in order to see whether the distribution of data influence the result. There are 5 text length features for a business and 5 text length feature for a user.

All the text length features are numbers. The text length features of the user and business might include the information of the writing style and informativeness of the reviews. The 5 metrics might display some potential relationship between the activity of the account and the usefulness. Comparing the text length of review, text length of user and text length of business might be helpful in the usefulness prediction.

Capital letter features(10)

The capital letter features can be divided into 2 branches: the number of capital letter of the business's all reviews and the number of capital letter of the user's all reviews. The capital letter features here are correlated with the user and business, so it's different with the capital letter feature in the basic feature section. We calculate the average, median, minimum, maximum and standard deviation value of the capital letter features in order to see whether the distribution of data influence the result. There are 5 capital letter features for a business and 5 capital letter feature for a user.

All the capital letter features are numbers. The capital letter features of the user and business might include the information of the number of important words. The 5 metrics might display some potential relationship between the activity of the account and the usefulness. Comparing the capital letter of review, capital letter of user and capital letter of business might be helpful in the usefulness prediction.

Punctuation count features(10)

The punctuation count features can be divided into 2 branches: the number of punctuation of the business's all reviews and the number of punctuation of the user's all reviews. The punctuation count features here are correlated with the user and business, so it's different with the punctuation count feature in the basic feature section. We calculate the average, median, minimum, maximum and standard deviation value of the punctuation count features in order to see whether the distribution of data influence the result. There are 5 punctuation count features for a business and 5 capital letter feature for a user.

All the punctuation count features are numbers. The punctuation count features of the user and business might include the information about clarity and writing style. The 5 metrics might display some potential relationship between the activity of the account and the usefulness. Comparing the punctuation count of review, punctuation count of user and punctuation count of business might be helpful in the usefulness prediction.

Rate features(10)

The rate features can be divided into 2 branches: the rate of the business's all reviews and the rate of the user's all reviews. The rate features here are correlated with the user and business, so it's different with the rate feature in the basic feature section. We calculate the average, median, minimum, maximum and standard deviation value of the rate features in order to see whether the distribution of rate influence the result. There are 5 rate features for a business and 5 rate feature for a user.

All the rate features are numbers. Most of the time the system only provide the average rate of business for the user. We are curious to see if there is any relationship between rate distribution features and review helpfulness. The 5 metrics might display some potential relationship between the activity of the account and the usefulness. We are interested in figuring out the rate giving style of the user and the rate receiving pattern of the business. Are the user "sweet" or "bitter" reviewers? 4 star might be a praise for one user and be a blame for other user. Comparing the rate of review, rate of user and rate of business might be helpful in the usefulness prediction.

The above features capture the textual and token-based characteristics of the user behaviour and business performance. We hypothesize that longer reviews may contain more information about the products and the reviewers might find useful. A long review tends to have more words, punctuation and capital letters. The 5 metrics are focus on the summation of daily activity, trying to diminished the random error, especially the influence of auto review generator or the biased reviewers.

4.2 Model

Our model is a stack ensemble model because according to the experience ensemble methods are commonly used to boost predictive accuracy by combining the predictions of multiple machine learning models and it works best in most of the cases. The models used in the ensemble model are Gradient Boosting Models (GBM) and Random Forests (RF). GBM and RF are both tree models, the advantage of tree models are 1. fast to train 2. get the list of feature importance easily 3. the logic behind the model hides in the layout of the trees and easy to understand. We know that error = bias + variance, the error is a trade-off between bias and variance. The GBM is based on weak learners with high bias and low variance(under-fit) while the RF is based on fully grown decision trees with low bias and high variance(over-fit).

Stacking (sometimes called stacked generalization) involves training a learning algorithm to combine the predictions of several other learning algorithms. First, all of the other algorithms are trained using the available data, then a combiner algorithm is trained to make a final prediction using all the predictions of the other algorithms as additional inputs. If an arbitrary combiner algorithm is used, then stacking can theoretically represent any of the ensemble techniques described in this article, although, in practice, a logistic regression model is often used as the combiner.

Stacking typically yields performance better than any single one of the trained models (Wolpert, 1992). It has been successfully used on both supervised learning tasks (regression(Breiman, 1996), classification and distance learning (Ozay, Yarman, 2013)) and unsupervised learning (density estimation)(Smyth, Wolpert, 1999). It has also been used to estimate bagging's error rate (Wolpert, Macready, 1999). It has been reported to out-perform Bayesian model-averaging (Clarke, 2003). The two top-performers in the Netflix competition utilized blending, which may be considered to be a form of stacking(Sill, Takacs, Mackey, Lin, 2009).

We use the framework call the H2O.ai to build the models. H2O provide stacked ensembles methods to use multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms.

V. Result

We trained and evaluated models using 5-fold cross-validation and report average performance on all five held-out folds. The reported accuracy is divided into training accuracy and testing accuracy. While creating the dataset for training and testing, we used the threshold of ζ = 2 in deciding whether a review is useful. Reviews with two or more "usefulness votes" were considered useful and reviews with less than two "usefulness votes" were considered not useful. Levi & Mokryn (2014) used Yelp data and considered a threshold of ζ = 5.

This dataset include 37% usefulness reviews and 63% unuseful reviews. The formula for the accuracy is:

A = (TP + TN)/(TP + FN + FP + TN)

TP: true positive

TN: true negative

FP: false positive

FN: false negative

and we get the testing accuracy is 82.5%. The formula for the recall is:

R = TP/(TP+FN)

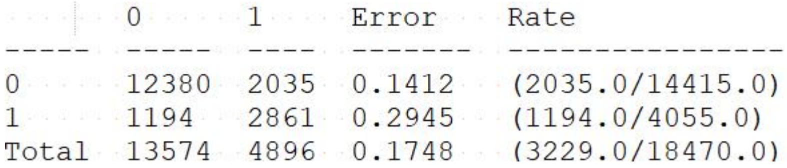
and we get the testing recall is 58.4%. The formula for the precision is:

P = TP/(TP+FP)

and we get the precision is 70.6%. The formula for the f-measure is:

F = 2*R*P/(R+P)

The f-measure is 0.639. The ensemble test AUC is 92.8% from the report of the model.



```
        0      1     Error    Rate
-----  -----  ----  -------  ----------------
0      12380  2035  0.1412   (2035.0/14415.0)
1      1194   2861  0.2945   (1194.0/4055.0)
Total  13574  4896  0.1748   (3229.0/18470.0)
```

Table 2: Testing Confusion Matrix (Act/Pred)

The top 12 variable importance are: rate_of_business_max, rate_of_business_sd, word_count, rate, vote_of_user_sd, number_of_cap_review, punctuation_count, user_review_date_sd, user_text_len_sd, business_day_average, user_text_len_max and punctuation_user_sd.

From this plot we find that rate_of_business_max, rate_of_business_sd, word_count, rate, vote_of_user_sd, number_of_cap_review are the most important features in this model.

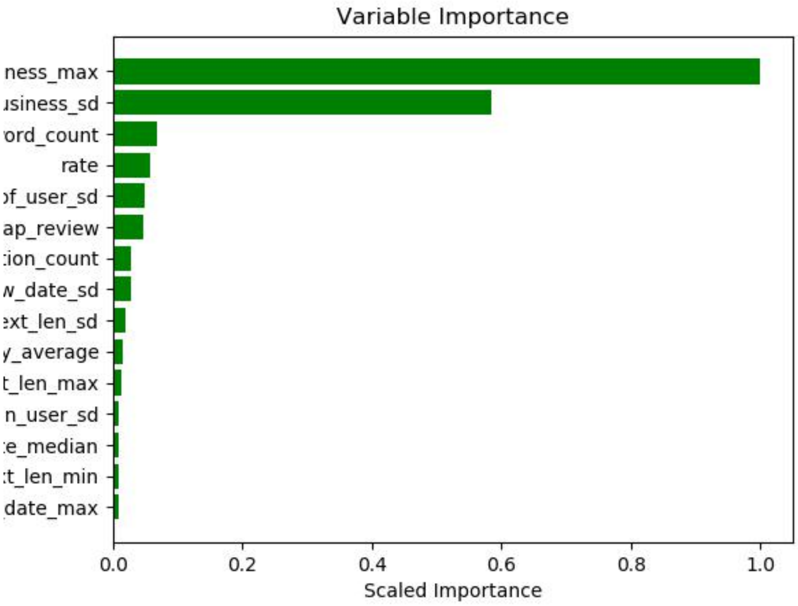


Figure 4: The variable importance of top 15 variables

The scoring history(training_classification_error and training_auc) of each iteration of the model is list as follow:

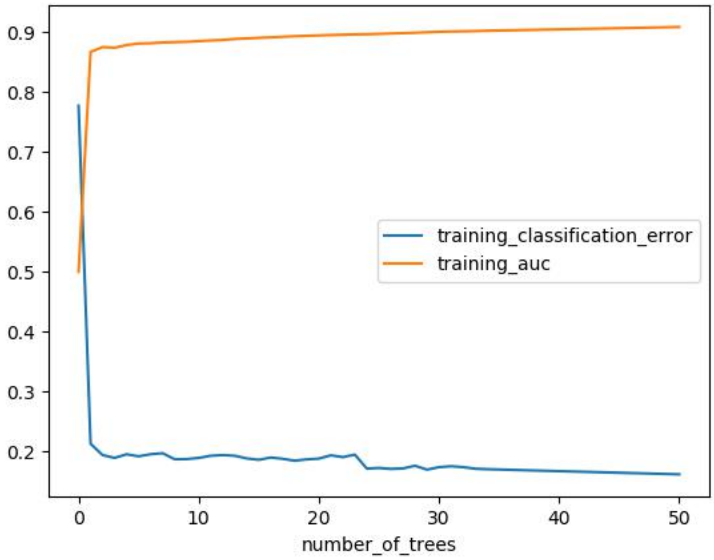


Figure 5: The relation between the training_classification_error, training_auc and the number of trees

The training classification error decrease dramatically after the number of trees increase to 3. The training classification then slight decrease as the number of trees increase.

After the feature importance research, we use the top 12 feature in sequence to do the feature ablation analysis. The result is shown in the chart(The accuracy of the model is 82.5%):

features	percent change	accuracy
rate_of_business_max	-4.24%	79.02%
rate_of_business_sd	-1.73%	81.09%
word_count	-0.63%	82.00%
rate	-1.10%	81.61%
vote_of_user_sd	-1.79%	81.04%
number_of_cap_review	-1.45%	81.32%
punctuation_count	-1.51%	81.27%
user_review_date_sd	-1.67%	81.14%
user_text_len_sd	-0.71%	81.93%
business_day_average	-0.08%	82.45%

user_text_len_max	-1.51%	81.27%
punctuation_user_sd	-0.71%	81.93%

Table 3. Accuracy and percent change of models trained using all but one feature type. A large drop in performance indicates the marginal contribution of the feature type

From this chart we find that the max rate of business is the most influential features of the dataset. All the other features are not effective for the accuracy change. Part of the reason might be the number of the features because there are 60 features in this model.

VI. Discussion and Conclusion

The results of our study show that the this model can effectively predict the usefulness of online catering industry reviews marked by users. The binary classifiers showed fairly good performance (0.825 when $\zeta = 2$). The good performance of the binary classifiers is most likely because the high-quality features is include and the performance of the stack ensemble model.

The basic features have a great performance in this studies. The rate and word count of the review have great impact on the review usefulness compared with most of the other variables. The more characters, words, and sentences one review has, the more usefulness it is. This finding support the H1. The text length report of the user and business are not as important as the word count of the review. The number of capital letter and the punctuation count also have positive impact on the model.

The sentiment features work bad in this studies. The influence of review sentiment is not as great as it was mention in the past studies. Part of the reason is related to the different tools used in the studies or it is not suitable for the catering category.

The date history feature is not useful in this model. I think the reason might relate to the audience's reading behaviour. Most of the audiences don't check the activities of the user or business. They are more focus on the review itself than the source history activities.

The max text length of the user and the standard deviation of the text length of the user is an interesting feature relate with the usefulness. This might show the relation between the usefulness and the user history activities. There might be 2 reasons for the 2 metrics: first, these reviewers are unbiased, each of their reviews fully describe the products/services of the business. The length of each of their reviews is not change dramatically but steadily. The standard deviation of the text length is keep in a rational range but large than zero. The max text length might relate with the word count in basic features. Most of the time more words means more information.

Most of the features don't have great impact on the performance of the model. These features are not commonly used to prioritize reviews for users. Our study points that several uncommon features and metrics especially the standard deviation could be used by online recommender systems to rank and display reviews for users and expose the content that is more often perceived to be useful to the user.

Overall the binary models show fairly high performance even the threshold is low(0.825 when $\zeta = 2$).

VII. Future Work

This model can be extended to other category of business such as hotel and bar. We believe that by limiting the category to catering, the model gains some advantages because there could be more noises caused by the difference between various domains.

Though we try a lot of features in this studies, only part of them contribute significantly to performance. More features should be considered in the future research, such as the influence of the reviewer’s social network and the readability. The standard deviation prove that this metric works good in the model. Standard deviation might relate to the activities of the user or the business. The research of the data change or activities report in a given time period might be useful in the further study.

The usefulness vote might relate to the number of reviewers who see this comment. A useful review could receive 0 useful vote and be seem as unuseful if no one read this review. The number of reviewers who read the review is not include in the dataset, maybe in the future this features is include and can be used in the further study.

Acknowledgment

The completion of this report could not have been possible without the guidance of Professor Jaime Arguello and Mr. Heejun Kim.

Thank you sincerely.

Reference

Levi, A., & Mokryn, O. (2014). The social aspect of voting for useful reviews. *Paper presented at the International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction,* pp. 293-300.

Chu, W., & Roh, M. (2014). Exploring the role of preference heterogeneity and causal attribution in online ratings dynamics. *Asia Marketing Journal*, 15(4), 61-101.

Forman, C., Ghose, A., & Wiesenfeld, B. (2008). Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets. *Information Systems Research*, 19(3), 291-313.

Hu, N., Liu, L., & Zhang, J. (2008). Do online reviews affect product sales? The role of reviewer characteristics and temporal effects. *Information Technology Management*, 9(3), 201-214.

Cao, Q., Duan, W., & Gan, Q. (2011). Exploring determinants of voting for the ‘‘helpfulness’’ of online user reviews: a text mining approach. *Decision Support Systems*, 50, 511-521.

Racherla, P., & Friske, W. (2012). Perceived ‘usefulness’ of online consumer reviews: An exploratory investigation across three services categories. *Electronic Commerce Research and Applications*, 11(6), 548-559.

Baek, H., Ahn, J. H., & Choi, Y.-S. (2012). Helpfulness of online consumer reviews: Readers’ objectives and review cues. *International Journal of Electronic Commerce*, 17(2), 99-126.

Li, M., Huang, L., Tan, C.-H., & Wei, K.-K. (2013). Helpfulness of online product reviews as seen by consumers: Source and content features. *International Journal of Electronic Commerce*, 17(4), 101-136.

Qing, C., Wenjing, D., & Qiwei, G. (2011). Exploring determinants of voting for the “helpfulness” of online user reviews: A text mining approach. *Decision Support Systems*, 511-521.

Biswas, D., & Biswas, A. (2004). The diagnostic role of signals in the context of perceived risks in online shopping: Do signals matter more on the web? *Journal of Interactive Marketing*, 18(3), 30-45.

Rieh, S. Y., & Belkin, N. (2000). Interaction on the web: Scholars' judgment of information quality and cognitive authority. *Proceedings of the Annual Meeting-American Society for Information Science*, 37. pp. 25-38.

Park, D.-H., Lee, J., & Han, I. (2007). The Effect of On-Line Consumer Reviews on Consumer Purchasing Intention: The Moderating Role of Involvement. *International Journal of Electronic Commerce*, 125-148.

Clemons, E. K., Gao, G. G., & Hitt, L. M. (2006). When online reviews meet hyperdifferentiation: A study of the craft beer industry. *Journal of Management Information Systems*, 23(2), 149-171.

Jackson, T. W., & Farzaneh, P. (2012). Theory-based model of factors affecting information overload. *International Journal of Information Management*, 32(6), 523-532.

Jacoby, J. (1977). Information load and decision quality: Some contested issues. *Journal of Marketing Research*, 14, 569-573.

McCormick, E. (1970). Human Factors Engineering. *New York: McGraw-Hill Book Company.*

Hildon, Z., Allwood, D., & Black, N. (2012). Impact of format and content of visual display of data on comprehension, choice and preference: A systematic review. *International Journal for Quality in Health Care*, 24(1), 55-64.

Lipowski, Z. (1975). Sensory and information inputs overload. *Comprehensive Psychiatry*, 16(3), 199-221.

Driver, M., & Mock, T. (1975). Human information processing decision style theory, and accounting information systems. *Accounting Review, 50(3)*, 490-508.

O’Reilly (1980). Individuals and information overload in organisations: Is more necessarily better? *Academy of Management Journal, 23(4)*, 684-696.

Jakoby, J., Speller, D., & Kohn, C. (1974). Brand choice as a function of information load. *Journal of Marketing Research, 11(1)*, 63-69.

Schultz, C., Schreyoegg, J., & von Reitzenstein, C. (2013). The moderating role of internal and external resources on the performance effect of multitasking: Evidence from the R&D performance of surgeons. *Research Policy, 42(8),* 1356-1365.

Van Rooy, R. (2003). Quality and quantity of information exchange. *Journal of Logic, Language and Information, 12(4),* 423-451.

Ganu, G., Kakodkar, Y., & Marian, A. (2013). Improving the quality of predictions using textual information in online user reviews. *Information Systems, 38(1),* 1-15.

Grice, H. (1967). Logic and conversation. *In P. Grice (Ed.), Studies in the way of Worlds. Cambridge, MA: Harvard University Press.*

Dellarocas, C. (2003). The digitization of word of mouth: Promise and challenges of online feedback mechanisms. *Management Science: Special Issue on E-Business and Management, 49(10),* 1407-1424.

Li, J., & Zhan, L. (2011). Online persuasion: How the written word drives WOMevidence from consumer-generated product reviews. *Journal of Advertising Research-New York, 51(1)*, 239-257.

Sotiriadis, M. D., & van Zyl, C. (2013). Electronic word-of-mouth and online reviews in tourism services: The use of twitter by tourists. *Electronic Commerce Research, 13(1),* 103-124.

Anderson, M., & Magruder, J. (2012). Learning from the crowd: Regression discontinuity estimates of the effects of an online review database. *The Economic Journal, 122(563)*, 957-989.

Luca, M., & Zervas, G. (2013). Fake it till you make it: Reputation, competition, and yelp review fraud. *Harvard Business School NOM Unit Working Paper, (14-006)*

Korfiatisa, N., García-Bariocanalb, E., & Sánchez-Alonso, S. (2012). Evaluating content quality and helpfulness of online product reviews: The interplay of review helpfulness vs. review conten. *Electronic Commerce Research and Applications,* 205-217.

Huang, A. H., Chen, K., Yen, D. C., & Tran, T. P. (2015). A study of factors that contribute to online review helpfulness. *Computers in Human Behavior, 48,* 17-27.

Keller, K., & Staelin, R. (1987). Effects of quality and quantity of information on decision effectiveness. *Journal of Consumer Research, 14(2)*, 200-213.

Alkhattabi, M., Neagu, D., & Cullen, A. (2011). Assessing information quality of elearning systems: A web mining approach. *Computers in Human Behavior,* 27, 862-873.

Arazy, O., & Kopak, R. (2011). On the measurability of information quality. *Journal of American Society of Information Sciences and Technology, 62(1)*, 89-99.

Leung, H. (2001). Quality metrics for intranet application. *Information & Management, 38(3)*, 37-152.

Wang, R., & Strong, D. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems, 12(4),* 5-33.

Yaari, E., Baruchson-Arbib, S., & Bar-Ilan, J. (2011). Information quality assessment of community generated content: A user study of Wikipedia. *Journal of Information Science, 37(5),* 487-498.

Chevalier, J., & Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research, 43(3),* 345-354.

Gupta, P., & Harris, J. (2010). How e-WOM recommendations influence product consideration and quality of choice: A motivation to process information perspective. *Journal of Business Research, 63(9),* 1041-1049.

Yang, J., Kim, W., Amblee, N., & Jeong, J. (2012). The heterogeneous effect of WOM on product sales: Why the effect of WOM valence is mixed? *European Journal of Marketing, 46(11/12),* 1523-1538.

Mudambi, S., & Schuff, D. (2010). What makes a helpful online review? A study of customer reviews on Amazon.com. *MIS Quarterly, 34(1),* 185-200.

Poston, R., & Speier, C. (2005). Effective use of knowledge management systems: A process model of content ratings and credibility indicators. *MIS Quarterly, 29(2),* 221-244.

Ahluwalia, R., Burnkrant, R., & Unnava, R. (2000). Consumer response to negative publicity: The moderating role of commitment. *Journal of Marketing Research, 37,* 203-221.

Hao, Y., Li, Y., & Zou, P. (2009). Why some online product reviews have no usefulness rating? *In Proceedings of the 2009 Pacific Asia conference on information systems.*

Daft, R. L., & Lengel, R. H. (1986). Organizational information requirements, media richness and structural design. *Management Science, 32(5)*, 554-571.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment Classification using Machine Learning. *ACL-02 conference onempirical methods in natural language processing (pp. 79-86).* Philadelphia: ACL.

Liu, Y., Huang, X., An, A., & Yu, X. (2008). Modeling and Predicting the Helpfulness of Online Reviews. *2008 Eighth IEEE International Conference on Data Mining (pp. 443 - 452). IEEE.*

Cheung, M. Y., Luo, C., Sia, C. L., & Chen, H. (2009). Credibility of electronic word-of-mouth: Informational and normative determinants of on-line consumer recommendations. *International Journal of Electronic Commerce, 13(4)*, 9-38.

Schlosser, A. E. (2005). Posting versus lurking: Communicating in a multiple audience context. *Journal of Consumer Research, 32(2),* 260-265.

Sweeney, J. C., Soutar, G. N., & Mazzarol, T. (2008). Factors influencing word of mouth effectiveness: Receiver perspectives. *European Journal of Marketing, 42(3/4),* 344-364.

Grewal, D., Gotlieb, J., & Marmorstein, H. (1994). The moderating effects of message framing and source credibility on the price-perceived risk relationship. *Journal of*

Consumer Research, 21(1), 145-153.

Kanouse, D. (1984). Explaining negativity biases in evaluation and choice behavior: *Theory and research. Advances in Consumer Research, 11(1),* 703-708.

Hastak, M., & Park, J. (1990). Mediators of message sidedness effects on cognitive structure for involved and uninvolved audiences. *Advances in Consumer Research, 17(1),* 329-336.

Crowley, A., & Hoyer, W. (1994). An integrative framework for understanding twosided persuasion. *Journal of Consumer Research, 20(4),* 561-574.

Pan, Y., & Zhang, J. Q. (2011). Born Unequal: A Study of the Helpfulness of User-Generated Product Reviews. *Journal of Retailing,* 598-612.

Kim, S.-M., Pantel, P., Chklovski, T., & Pennacchiotti, M. (2006). Automatically assessing review helpfulness. *Conference on Empirical Methods in Natural Language Processing* (pp. 423-430). Sydney, Australia: Association for Computational Linguistics.

Opitz, D.; Maclin, R. (1999). "Popular ensemble methods: An empirical study". *Journal of Artificial Intelligence Research. 11*: 169-198.

Seni, G., & Elder, J. F. (2010). Ensemble methods in data mining: improving accuracy through combining predictions. *Synthesis Lectures on Data Mining and Knowledge Discovery, 2(1),* 1-126.

Kuncheva, L. and Whitaker, C., Measures of diversity in classifier ensembles, *Machine Learning, 51,* pp. 181-207, 2003