Classifying Business Marketing Messages on Facebook

Bei Yu
School of Information Studies
Syracuse University
1-315-443-3614
byu@syr.edu

Linchi Kwok
David B. Falk College of Sport and Human Dynamics
Syracuse University
1-315-443-2612
lkwok@syr.edu

ABSTRACT
Companies are increasingly using social media for marketing purposes. In this study we first demonstrate that although the majority of company posts on Facebook are aimed for direct sales and promotions, it is their communication messages that received the most attention from customers. We then trained an SVM classifier to automatically separate these two kinds of messages, hoping to use this tool to analyze messages from many companies and consequently monitor the evolution of their social media use over time. We found that the classifier trained with tf-idf weighted part-of-speech features performed best. It is better than classifiers trained with word features. Combining feature sets did not improve the performance. Feature ranking results show that this best-performed classifier captured the genre characteristics of direct marketing and communication messages.

Categories and Subject Descriptors
H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing – *linguistic processing.*

General Terms
Algorithms, Measurement, Design, Experimentation.

Keywords
Social media, advertisement, marketing, communication, machine learning, feature selection, text classification, text categorization.

1. INTRODUCTION
Social media can have significant impacts on consumer behaviors, and companies are increasingly using social media for marketing purposes [12]. Facebook has become one of the most dominant media for B2C (business-to-cusumer) and C2C (consumer-to-consumer) communications [14][16]. Today, many websites embed Facebook's "Like" button. Popularity of a company's Facebook messages could be important in indicating the effectiveness of the company's social media strategies.

While companies and scholars are finding ways to make good use of social-media tools [6], so far relatively few studies examine companies' social media messages sent on Facebook. Being able to identify the type of messages sent on Facebook, and measure the attention of these messages received from their Facebook fans, may assist managers in developing the "right" social-media marketing strategies.

In this study we first manually constructed a new typology to categorize the Facebook messages because existing typologies created for traditional media (e.g. TV and radio) cannot accommodate the new characteristics of social media, such as high interactivity. We then measured the popularity of messages in each category. We demonstrate that although the majority of the company posts aimed for direct sell and promotion, it is their communication messages that received the most attention from customers.

We then trained a text classifier based on restaurant messages to automatically separate these two kinds of messages, hoping to use this tool to analyze messages from many more companies and consequently monitor the evolution of their social media use over time. Direct marketing and communication messages have different communication goals, and thus can be considered as two different genres [19]. These two message types also differ in the topics they carry: marketing messages often highlight products, announce campaigns and boast achievements. Communication messages often deliver seasonal greetings and daily life suggestions, or provoke feedback. Therefore this classification task crosses genre classification and topic classification. We explored different feature sets that fit for genre- and topic-classification, and compare their performance.

This paper is organized as follows. Section 2 introduces the data scraping and pre-processing. Section 3 describes the new message typology and the relationship between message types and their popularity. Section 4 describes the classification experiment setup, including the training and testing data construction, classification algorithm, performance measure, and document representation. Section 5 describes the experiment results. Sections 6 and 7 conclude with discussions and future work.

2. DATA PREPARATION
We choose the restaurant industry as a case study because many restaurants have set up company pages on Facebook and attracted large number of fans [21]. We selected twelve restaurants according to the data sample procedure described in [23], downloaded the company posts several times between October and December 2010, and finally merged the downloaded batches into one data set consisting of 982 messages.

3. MESSAGE TYPES

3.1 The gold standard of message types
One hospitality expert used grounded theory and open coding method to code all messages and eventually constructed a new

typology of company messages on Facebook. Detailed description of the typology construction process can be found in [12]. The final typology including two tier-1 categories: marketing and communication. A marketing message is defined as a one-way and often persuasive message of selling or promoting a service, a product, or the brand to Facebook users. For example *"Carmine's Gift Cards make the best Holiday gift."* A communication message is defined as a one- or two-way message without direct sell or promotion information. For example "*Happy Father's Day!*."

The sample set includes two subsets, one downloaded from October and November, and one from December. The coder finished coding the first and then the second subset, not knowing that these two subsets actually overlapped with 200 identical messages appeared in both subsets. Therefore these messages were coded twice at different times without the coder's awareness. Thus the intra-coder agreement truly reflects the code reliability. According to the result in Table 1, the raw intra-coder agreement is 87.5%. Cohen's Kappa, a more strict measure, is 0.69, indicating less perfect but solid agreement [4]. As for the 25 disagreed cases, the second-time code served as the gold standard.

Table 1. Intra-coder agreement ("M" for Marketing and "C" for Communication)

Message type	M	C	total
M	132	10	142
C	15	43	58
total	147	53	200

3.2 Message types and popularity

We use the number of "like" responses and the number of comments to measure and compare the popularity of marketing and communication messages. The raw numbers cannot be directly used because the number of fans varies significantly between companies. Starbucks has over 14 million fans as of September 2010, but Carmines has only five thousands. A massage that attracts one thousand "like" responses should be considered very popular for Carmines but not so for Starbucks. We used z-scores to normalize the numbers of "like" responses and comments within each company so that the numbers are comparable across companies. If a message was just posted when we downloaded it, the number of "like" responses and comments did not truly reflect its ultimate popularity. To solve this problem, we downloaded data periodically (usually every two weeks) and then merged the downloaded batches, replacing the old numbers with new ones if a message occurred in two consecutively downloaded batches, and excluding the newest posts in the most recently downloaded batch. The current downloading frequency suffices because the restaurants do not post very often. We found a large proportion of overlapping messages between two consecutively downloaded batches.

Table 2 shows that marketing messages account for 73.3% of all messages, but communication messages received much more attention than direct marketing ones. This finding suggests that companies are utilizing the interactivity of social media to enhance customer communication, but they still spend most effort in "broadcasting" direct marketing information just as they use traditional media. This trend may change in the future when the companies become adept in engaging customers in social media. Our goal is to automatically monitor such evolution: download the data periodically, build automatic classifiers to separate marketing and communication messages, and eventually conduct longitudinal study on the trend of social media use.

Table 2. Message types and popularity

Message type	Avg. #Like (std.dev)	Avg. #Comments (std.dev)
M (720)	-.07 (.88)	-.14 (.76)
C (262)	.14 (1.19)	.36 (1.41)

4. EXPERIMENT SETUP

4.1 Splitting training and test data

It is likely that some companies post unique content (e.g. product names) in their messages, and thus when messages from the same company are included in both training and test sets, a classifier might pick up the unique characteristics and achieve deceptively high performance. To evaluate the classifier's performance across companies, we made sure the training and the test sets include messages from different companies.

Our sample set accumulates messages posted by twelve restaurants, including seven quick service (e.g. McDonald's), three casual dining (e.g. Olive Garden), and two independents (e.g. Carmine's). We choose four restaurants, including two quick service (Chick-fil-A and Dunkin's Donuts), one casual dining (Chili's Grill and Bar), and one independent (Carmine's) as the test set, and the remaining eight restaurants as training set. The restaurants in the test set were chosen by alphabetic order in each restaurant category.

The message type distribution in Table 3 shows that the categories are highly skewed, which poses a particular challenge for inductive learning [7]. Common solutions include introducing a special loss function to penalize prediction error in a certain category (e.g. mis-classifying a regular email as spam), or conducting instance sampling to reduce the number of examples in over-populated category [15].

In our problem there is no need to penalize either kind of prediction errors, hence we adopt a convenient instance sampling approach to reduce the number of marketing messages in the training set. We sorted all marketing messages in each company by their post IDs, which are automatically assigned by Facebook in chronological order, and then selected all marketing messages ranked in odd number. This approach reduced the number of marketing messages by half, and resulted in a nearly balanced training set with the marketing category accounting for 53.2% messages.

Table 3. Message type distribution

Data set	M	C	total	Majority
Train	483	213	696	69.4%

Test	237	49	286	82.9%
Total	720	262	982	73.3%
Balanced train	242	213	455	53.2%

4.2 Prediction performance measure

Because the test set also yields high class skew, the majority vote baseline is as high as 82.9%, but uesless. Hence accuracy is not an appropriate measure in this case [7]. Instead, we used a macro-averaged *F*-measure, which averages the F-measure in each category, to evaluate classifier performance [20].

4.3 Classification algorithm

SVMs (Support Vector Machines) are one of the best text classification methods [5][20] and feature selection methods [7][8]. SVMs select discriminative features with broad document coverage to reduce the risk of over-fitting [22]. In this study we used the SVM-light package [9] with default parameter settings.

4.4 Feature representation

Because the marketing messages focus on sales and promotions while the communication messages focus on personal interaction, we conjecture that these two types of messages differ in topics as well as language styles. With regard to topic, we found during the coding process that marketing messages involve highlighting products, announcing the beginning and result of marketing campaigns, boasting achievements and social responsibilities. Communication messages involve seasonal updates, daily life advice and suggestions, provoke feedback and call for action. With regard to language style, communication messages encourage two-way interactions and thus are more likely to use an engaging and involved style [2][3][1]. In comparison, marketing messages are one directional and thus less engaging.

Because this classification task crosses genre- and topic-classification, we utilize both genre-based language features and topic-based features to train the classifier, and compare their effectiveness in prediction. According to literature in genre classification [1][2][3][10][11][18], parts-of-speech distribution is a strong indicator of genre and style difference. For example, engaging and involved style is characterized by high percentage of pronouns, and less engaging and un-involved style by more articles and nouns. We use the part-of-speech tagger and shallow parser in the OpenNLP toolkit to process the messages and compute the frequency of each kind of part-of-speech and phrase.

We use Bag-of-Words (BOW) features for topic-based classification. The feature set includes all tokens tagged at word level by the OpenNLP toolkit. This means a word with multiple parts-of-speech will be treated as multiple tokens. To reduce the number of features, we removed all words that are used by only one company because they do not represent common topics.

After converting the messages to feature vectors, we also compare the effectiveness of four different feature representations: presence/absence (SVM-BOOL), frequency (SVM-TF), normalized frequency (SVM-NTF), and tf-idf (SVM-TFIDF). The tf-idf weighting of part-of-speech features follows the same formula as word tf-idf weighting, meaning the weight of a part-of-speech will be penalized if it is used in many messages.

5. RESULTS

5.1 Genre-based classification

We used 44 parts-of-speech as genre-based features to train the classifier. The prediction result in Table 4 shows that all four feature representations resulted in similar accuracy, from 74% to 76%, based on "leave-one-out" cross validation (abbreviated as "loo") on the training set, much better than the 53% majority baseline. On the test set, tf-idf representation yields the best performance. Its accuracy (83%) is just the same as the majority vote, but its macro-averaged *F* value is 0.67, higher than the majority vote baseline (0.50).

We further examined the feature ranking provided by the SVM-TFIDF classifier. The top five marketing indicators are number (CD), "to" (TO), DT (determiner), interjection (INTJ), and preposition or subordinate conjunction (IN). The top five communication indicators are wh-pronoun (WP), wh-adverb (WRB), non-3rd person singular present verb (VBP), superlative adjective (JJS), and modal (MD).

The SVM feature ranking result is consistent with previous findings in genre classification. The use of numbers (CD), determiners (DT) and interjections (INTJ) matches with the need of direct marketing messages to announce marketing campaign, highlight certain products, or boast achievements. The use of wh-words (WP and WRB), modals (MD), and superlative adjectives (JJS) characterizes the engaging and personalized style in communication messages.

Here is one example of marketing message using interjections.

"1600 Carmine's cookbooks sold on QVC Tuesday night in 7 minutes .. wow"

Here is one example of marketing message using determiners and numbers:

"Sunday's 12 Days of Sharing item is a Frosted Starbucks Tumbler for $6.99. Save a paper cup and keep your drink (and paws) warm."

Here is one communication message using wh- word:

"What is your wish for the holidays? Share it with us: http://starbucks.com/share"

Here is one communication message using modal:

"Now is a great time to register your Starbucks card. Did you know that you can do it from within Facebook?"

Table 4. Genre-based classification result

	train	Test			
algorithm	loo	Acc	F_M	F_C	F_{avg}
SVM-BOOL	.76	.74	.83	.38	.61
SVM-TF	.75	.74	.84	.40	.62
SVM-NTF	.73	.78	.87	.35	.61
SVM-TFIDF	.74	.83	.90	.44	**.67**

5.2 Topic-based classification

We used Bag-Of-Words features for topic-based classification. We removed the words that were used by only one company, no matter how many times they were used. Because function words and word forms can be informative features for text classification, we chose not to remove stop words or perform word stemming [16][22]. The final feature set includes 817 words. The classification result in Table 5 shows that all four feature representations reached near 80% accuracy by leave-one-out cross validation on the training set, much better than 53% majority baseline. On the test set, all classifiers reached 0.60 or higher macro-averaged *F* values. SVM-NTF and SVM-TFIDF tied with highest F value 0.64.

Table 5. Topic-based classification result

	train	test			
algorithm	Loo	Acc	F_M	F_C	F_{avg}
SVM-BOOL	.79	.69	.79	.42	.60
SVM-TF	.80	.71	.80	.43	.61
SVM-NTF	.80	.76	.84	.45	**.64**
SVM-TFIDF	.78	.77	.86	.43	**.64**

Table 6. Features ranked by SVM-NTF topic classifier

	Top indicators
M	*with, a, for, to, thanks, all, 1, by, gift, special, and, new, http, chicken, be, !, get, week, or, win, available, year*
C	*are, team, coming, favorite, too, many, what, third, latest, her, who, up, I, he, way, forward, we, thanksgiving, his*

Table 7. Features ranked by SVM-TFIDF topic classifier

	Top indicators
M	*thanks, a, gift, for, to, with, 2, thank, new, win, special, proud, try, all, 10, winners, make, want, year, chicken*
C	*what, are,team, ?, latest, favorite, I, many, too, coming, ‘s, weekend, things, her, thanksgiving, behind, but, his*

5.3 Combining genre and topic features

Since both genre classifiers and topic classifiers still performed lower than human agreement level, would combining genre- and topic-based features improve the prediction performance? We combined the two feature sets and re-conducted the classification experiments. The result in Table 8 shows that the benefit is negligible: the combined feature set improved the macro-averaged *F* value for SVM-BOOL and SVM-NTF classifiers by 0.02~0.03, but does not improve the SVM-TF and SVM-TFIDF classifiers.

So far the best performance is produced by the SVM-TFIDF classifier trained on 44 part-of-speech features. Adding word features did not help. A possible explanation is that the word features do not contribute new information. The parts-of-speech of the top word features ranked by SVM-NTF and SVM-TFIDF classifiers (Tables 6 and 7) coincide with the top-ranked parts-of-speech identified by the SVM-TFIDF genre classifier, confirming that the two feature sets actually describe the documents in similar ways, and that the part-of-speech tags are more succinct than words as descriptors.

Table 8. Combining genre- and topic-based features

	train	test			
algorithm	Loo	Acc	F_M	F_C	F_{avg}
SVM-BOOL	.82	.74	.83	.45	.64
SVM-TF	.79	.73	.83	.42	.62
SVM-NTF	.80	.77	.86	.46	**.66**
SVM-TFIDF	.78	.78	.86	.46	**.66**

5.4 Message length

We noticed that on average marketing messages are significantly longer than communication ones. In the training set, the average message length is 132 characters, comparable to the maximum length of tweets. The average length is 169 for marketing and 91 for communication messages. A trivial classifier may classify a message to marketing if it is longer than the average length (132), and to communication if not, and still result in 70% accuracy on the training set, and 0.59 macro-averaged *F* value on the test set.

Since message length is such a strong indicator, we normalized the message length using min-max normalization, and then combined the message length as a feature with part-of-speech features to re-train the genre classifiers, however the performance was not improved.

Table 9. Message length

category	N	Min	Max	Mean	std.dev
M	242	17	374	169	72
C	213	11	275	91	60
Total	455	11	374	132	77

6. CONCLUSIONS AND DISCUSSIONS

In this paper we aim to develop an automatic classifier to separate Facebook messages posted by restaurants into marketing and communication types, hoping to use the classifier as a tool to monitor the evolution of companies’ social media use.

Using SVMs as the classification algorithm, we compared the performance of part-of-speech features, word features, and their combination in four representations: Boolean, raw frequency, normalized frequency, and tf-idf weighting. The classifier trained with tf-idf weighted part-of-speech features is so far the best classifier we have constructed. This classifier captured the main characteristics of communication messages in an engaging and involving style, and marketing messages in a less engaging and “broadcasting” style. This classifier achieved macro-averaged *F* value 0.67, significantly higher than the majority vote baseline, but is not yet comparable to human expert’s level of consistency. One big challenge comes from the short message length, which is comparable to the length of tweets. Previous genre classification tasks usually used large corpora with long documents, resulting in more precise estimation of the proportion of parts-of-speech and other linguistic features [2][3][10][11][1][18].

7. FUTURE WORK

Besides restaurants, many other service industries have also been engaging customers in social media. To make this classifier really useful in monitoring marketing strategies in social media, we need to evaluate the classifier's performance in various domains. We have been downloading data from hotels, and will evaluate the classifier's cross-domain prediction performance in the future. We will also evaluate the classifier's time-sensitivity by manually examining its prediction performance in newly downloaded data set.

8. ACKNLOWLEDGMENTS

Our thanks to the Harrah Hospitality Research Center Grant Award Program from University of Nevada, Las Vegas.

9. REFERENCES

[1] Argamon, S., Koppel, M., Fine, J., and Shimoni, A.R. (2003). Gender, genre, and writing style in informal written texts. *Text*, 23(3): 321-346

[2] Biber, D. (1988). *Variation across speech and writing*. Cambridge University Press.

[3] Biber, D. (1995). *Dimensions of Register Variation: A Cross-linguistic Comparison.* Cambridge University Press.

[4] Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20:37-46.

[5] Dumais, S., Platt, J., Heckerman, D., and M. Sahami. (1998). Inductive learning algorithms and representations for text categorization. *Proceedings of the 7th International Conference on Information and Knowledge Management (CIKM'98)*, 148-155.

[6] Finin, T., Joshi, A., Kolari, P., Java, A., Kale, A., & Karandikar, A. (2008). The information ecology of social media and online communities. *AI Magazine*, 29(3): 77-92.

[7] Forman, G. (2003). An extensive empirical study of feature selection metrics for text categorization. *Journal of Machine Learning Research*, 3:1289–1305.

[8] Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machiness. *Machine Learning*, 46(1-3): 389–422.

[9] Joachims, T. (1998). Text categorization with Support Vector Machines: Learning with many relevant features. *Lecture Notes in Computer Science (ECML'98)*, Issue 1398, 137-142.

[10] Karlgren, J., and D. Cutting (1994). Recognizing text genres with simple metrics using discriminant analysis. *Proceedings of COLING'94.*

[11] Kessler, B., Nunberg, G., and Schutze, H. (1997). Automatic detection of text genre. *Proceedings of ACL/EACL'97*, 32-38.

[12] Kwok, L. and Yu, B. (2011). Towards developing a typology of hospitality companies' Facebook messages. Manuscript submitted for publication, Syracuse University

[13] Mangold, W. G., & Faulds, D. J. (2009). Social media: The new hybrid element of the promotion mix. *Business Horizons*, *52*(4): 357-365

[14] Nelson, R. (2010). *Tech bytes* [Television series]. New York: ABC Network.

[15] Ng, V. and Cardie, C. (2002). Combining sample selection and error-driven pruning for machine learning of coreference rules. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'02)*, 55-62

[16] Riloff, E. (1995). Little words can make a big difference for text classification. *Proceedings of the 18th ACM International Conference on Research and Development in Information Retrieval (SIGIR'95)*, 130-136

[17] Rubin, C. (2010). Should you advertise on Facebook? *Inc.com*. Retrieved on March 25, 2010 from http://www.inc.com/news/articles/2010/03/facebook-tops-google-for-loyalty.html?partner=newsletter_Technology

[18] Stamatatos, E., Fakotakis, N., and Kokkinakis, G. (2000). Text genre detection using common word frequencies, *Proceedings of COLING'00*.

[19] Swales, J. (1990). *Genre Analysis: English in Academic and Research Setting.* Cambridge University Press.

[20] Yang, Y. & Liu, X. (1999). A re-evaluation of text categorization methods. Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development *in Information Retrieval (SIGIR'99)*, 42–49

[21] Young, L.(2010). Brave New World. *Foodservice and Hospitality*, 42(11): 24-28

[22] Yu, B. (2008). An evaluation of text classification methods for literary study. *Journal of Literary and Linguistic Computing*, 23(3): 327-343

[23] Yu, B., Chen, M., and Kwok, L. (2011). Toward Predicting Popularity of Social Marketing Messages. *Proceedings of the 2011 International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction (SBP11)*, 317-324