Toward Predicting Popularity of Social Marketing Messages

Bei Yu[1], Miao Chen[1], and Linchi Kwok[2]

[1] School of Information Studies
[2] College of Human Ecology, Syracuse University
{byu,mchen14,lkwok}@syr.edu

Abstract. Popularity of social marketing messages indicates the effectiveness of the corresponding marketing strategies. This research aims to discover the characteristics of social marketing messages that contribute to different level of popularity. Using messages posted by a sample of restaurants on Facebook as a case study, we measured the message popularity by the number of "likes" voted by fans, and examined the relationship between the message popularity and two properties of the messages: (1) content, and (2) media type. Combining a number of text mining and statistics methods, we have discovered some interesting patterns correlated to "more popular" and "less popular" social marketing messages. This work lays foundation for building computational models to predict the popularity of social marketing messages in the future.

Keywords: text categorization, marketing, social media, prediction, media type.

1 Introduction

Social media has become a major influential factor in consumer behaviors, such as increasing awareness, sharing information, forming opinions and attitudes, purchasing, and evaluating post-purchasing experience [8]. Among those social media tools that are available for businesses, Facebook has become one of the most important media for B2C and C2C communications. Today many websites embed Facebook's "Liked" buttons, which not only allows consumers to exchange information within their Facebook network in a much faster speed but also pushes companies to pay more attention to Facebook fans' reactions to their messages sent through social media. Facebook fans' endorsement of a company's messages could be an important indicator of how effectiveness a company's social marketing strategies are.

This research[1] aims to examine the effectiveness of different social marketing strategies. In social media such as Facebook, the long-term effectiveness of marketing strategies, such as the impact on consumer behavior, might not be directly measurable. However, Facebook provides some measures that indicate the message popularity, which is the level of attention that fans paid to the messages, such as the number of

[1] This research project is supported by the Harrah Hospitality Research Center Grant Award Program from University of Nevada, Las Vegas.

J. Salerno et al. (Eds.): SBP 2011, LNCS 6589, pp. 317–324, 2011.

"likes", the number of comments, and the overall tone of the comments. To some extent the popularity of a message indicates the effectiveness of the marketing strategy embedded in this message. Therefore in this research we aim to discover the characteristics of social marketing messages that affect their popularity. We focus on the number of "likes" they attract, and leave the comment analysis for future work.

Using the number of "likes" to measure the popularity of a message, we aim to answer the following research question: what characteristics of the messages contribute to their popularity? We examine the characteristics of messages from two perspectives: (1) content, and (2) media type. This article is organized as follows. Section 2 describes the data preparation process, including data sampling, downloading, and cleaning. Section 3 describes the thematic analysis of the message content. Section 4 describes an approach to predicting popularity based on the message content. Section 5 describes the correlation analysis between media type and message popularity. Section 6 concludes with limitations and future work.

2 Data Sampling, Downloading, and Cleaning

In this section we describe the data sampling, downloading, and cleaning process. We choose restaurant industry as a case study for the following reasons. Restaurant business has always been a big component of the tourism and hospitality industry, where intangible and perishable products are produced and consumed. Due to the experiential nature of these service products, consumers often seek information from online reviews and travel blogs before making purchasing decisions [10]. Restaurants understand the importance of communicating the "right" messages with their existing and potential customers over the Internet. Recent business reports also showed that restaurant entrepreneurs can substantially draw business and increase sales with Twitters and Facebook [11]. Many restaurants have been actively engaged with their customers on Facebook, which makes restaurant industry a great example for examining companies' social marketing messages.

This research adopted a criterion-based, mixed-purpose sampling procedure and selected twelve restaurants' Facebook pages in September 2010 based on the following criteria and steps: (1) we examined the top 20 restaurant chains with the highest sales volume [7] and selected all quick service restaurant chains with more than one million Facebook fans (eight in total) and the top three casual dining restaurant chains with the most Facebook fans; (2) we examined the top 20 independent restaurants and selected the top two with the most Facebook fans. We should have selected 13 restaurants in total, but eventually included 12 because KFC's Facebook page couldn't be correctly identified.

We then identified each restaurant's unique Facebook ID and retrieved all posts using Facebook FQL. We retrieved three fields: message body, message's media type ("status", "link", "video", or "photo"), and the number of people who have clicked the "likes" button on this message.

On October 5th we retrieved a total of 675 messages posted by the above 12 restaurants. See Table 1 for the number of posts collected for each restaurant. So far Facebook FQL only returns posts in recent 30 days or 50 posts, whichever is greater. We plan to continue our data collection on a regular basis for the next six months and re-examine our analysis on the larger data set in the future.

Because the number of fans for each restaurant varies significantly, messages from different restaurants with same number of "likes" could indicate significantly different level of popularity. For example, the most popular message from independent restaurant Carmines attracted only 93 likes, while the least popular message from Starbucks attracted 2,250 likes. To make the numbers of "likes" comparable among restaurants we normalized the numbers of "likes" for each restaurant using z-score. The normalized numbers of "likes" are distributed in the range (-3, 2).

Table 1. Descriptions of the 12 restaurants in this study

Restaurant Name	Facebook ID	No. of FB Fans	No. of FB posts
Quick Service (7)			
McDonald's	50245567013	3,244,673	39
Starbucks	22092443056	14,605,410	58
Subway	18708563279	2,218,422	58
Pizza Hut	6053772414	1,473,274	63
Taco Bell	18595834696	2,842,184	102
Dunkin's Donuts	6979393237	2,078,568	45
Chick-fil-A	21543405100	3,106,304	73
Casual Dining (3)			
Chilli's Grill & Bar	106027143213	470,994	51
Olive Garden	18618646804	544,278	41
Outback Steakhouse	48543634386	883,670	40
Independents (2)			
Joe's Stone Crab	404077308008	1,628	54
Carmine's	99057540626	5,190	51

3 Thematic Analysis

Identifying social marketing strategies is difficult without an established typology. However we may identify the common themes of the messages as indicators of restaurants' favorite marketing strategies. For example, if "donation" occurs frequently we can infer that charity is a commonly used marketing strategy. In order to identify the main themes we first applied Latent Dirichlet Allocation (LDA) algorithms [1] to all messages and identified ten main themes. However, the resulted themes were not as meaningful probably because LDA has proved to be more effective on larger data set. We then turned to basic keyword analysis to identify the most common content words (nouns, verbs, adjectives, and adverbs) that each restaurant used as a representation of its marketing strategies. Because of the important role of pronouns in communication [2], we also counted pronouns in this study.

During the keyword analysis process we (1) gathered all messages posted by each restaurant as an individual text collection; (2) tokenized the messages and computed each token's overall term frequency (how many times it occurs in this text collection); (3) manually picked the ten most frequent pronouns and content words that consist of the "keyword profile" of each restaurant. The resulted keyword profiles are presented in Table 2.

The keyword profiles in Table 2 depict a common communicative style signified by prevailing use of first and second person pronouns like “you”, “your”, “we”, and “our”. These profiles also uncover two common marketing strategies. The first is to promote key concepts as part of their public images. For example, McDonald’s promoted “family time”, Subway emphasized “fresh” food and being “fit”, and Starbucks focused on the concept of “enjoy”. The second common strategy is to run advertisement campaigns. Many of the keywords other than pronouns involve the major advertising campaigns, especially contests. Almost all restaurants launched some contests, for example “Super Delicious Ingredient Force” by TacoBell, “Ultimate Dream Giveaway” by Pizza Hut, “Create-A-Pepper” by Chili’s, “Never Ending Pasta Bowl” by OliveGarden, “Free Outback For a Year” by Outback, “Eat Mor Chickin Cowz” by Chick-fil-A, “Fan Of The Week” by Dunkin’s Donuts, etc.

Table 2. Keyword profiles for each restaurant

Restaurant	Keywords
Quick Service (7)	
McDonald’s	you, your, games, all, we, time, Olympic, angus, more, howdoyoumcnugget
Starbucks	you, we, your, coffee, our, free, us, about, enjoy, more
Subway	your, you, favorite, fresh, footlong, today, chicken, fit, buffalo, breakfast
Pizza Hut	your, pizza, you, our, today, we, order, free, new, dream
Taco Bell	you, your, who, force, night, sdif, fourthmeal, super, chicken, delicious
Dunkin’s Donuts	you, FOTW, our, coolatta, week, DD, next, your, bit, submit
Chick-fil-A	cowz, you, mor, your, cow, day, facebook, our, like, appreciation
Casual Dining (3)	
Chilli’s Grill & Bar	we, you, fans, our, your, pepper, us, winner, St. Jude, gift
Olive Garden	you, our, your, pasta, new, we, ending, bowl, never, contest
Outback Steakhouse	bit, free, your, our, year, you, win, check, visit, chance
Independents (2)	
Joe’s Stone Crab	special, week, our, dinner, crab, jazz, summer, we, you, sauce
Carmine’s	your, our, we, family, memories, my, Washington DC, us, you, dinner

Did the key concepts resonate among fans? Did these contests really catch the eyes of fans? Or more broadly which strategies are more effective? By measuring the effectiveness with the popularity of relevant messages, we translated this question to the question of which messages are more popular than others, and then adopted an exploratory approach of using text categorization and feature ranking methods to discover the relationship between message content and popularity.

4 Content-Based Message Popularity Prediction

We grouped all messages into two categories: the “more popular” category for messages with positive normalized numbers of “likes” (above average), and the “less popular” category for messages with negative normalized numbers of “likes” (below average). After this process, the messages are separated into two categories: 430 messages in the “less popular” category and 207 messages in the “more popular” category. To build a sharp and balanced classifier we formed a new data set “400 messages" which consists of 200 most popular and 200 least popular messages. Each message is tokenized and converted to a word vector using bag-of-word representation. To focus on common words we excluded words that occurred in fewer than three messages and resulted in a vocabulary of 568 words. Reducing the vocabulary size also resulted in 19 empty vectors in the "400 messages" data set, which eventually has 381 (192 "more popular", and 189 "less popular") messages.

We then apply text categorization algorithms to examine the extent to which the "more popular" and "less popular" categories are separable. We employed two widely used text categorization algorithms Support Vector Machines (SVMs) and naïve Bayes (NB) to train and evaluate the content-based popularity prediction models.

SVMs are among the best methods for text categorization [6] and feature selection [4]. SVMs favor discriminative features with broad document coverage, and therefore reduce the risk of over-fitting. Studies have shown that the 10% most discriminative features identified by SVM classifiers usually work as well as, or even better than the entire original feature set [12]. In this study we used the SVM-light package [5] with default parameter setting. We compared the performance of four variations of SVM models: SVM-bool with word presence or absence (one or zero) as feature values; SVM-tf with raw word frequencies; SVM-ntf with word frequencies normalized by document length; and SVM-tfidf with word frequencies normalized by document frequencies.

Naïve Bayes (NB) is another commonly used text categorization method, simple but effective even though the text data often violate its assumption that feature values are conditionally independent given the target value [3]. We compared two variations of naïve Bayes implementations: NB-bool, the multi-variate Bernoulli model that uses word presence and absence as feature values, and NB-tf, the multinomial model that uses raw word frequencies. We implemented the two naïve Bayes variations.

We used leave-one-out cross validation to evaluate the SVMs and NB classifiers on both "all messages" and "400 messages" data sets. We used the majority vote as the baseline for comparison. The results in Table 3 show that all classifiers beat the majority vote on the "400 messages" data set by wide margins, and that SVM-ntf and SVM-bool both achieved accuracy over 75%. In contrast the cross validation accuracies on the "all messages" data set are less encouraging in that they are merely comparable to the majority vote. The good result on "400 messages" demonstrates that the "more popular" and "less popular" messages are fairly separable based on message content. The poor result on "all messages" also supports this notion in that the messages with medium level of popularity proved to be difficult to categorize.

Table 3. Leave-one-out cross validation results (in percentage)

Algorithm	All messages	400 messages
Majority vote	67.5	50.4
Svm-bool	67.1	*75.1*
Svm-tf	67.8	74.3
Svm-ntf	72.2	*75.9*
Svm-tfidf	71.9	71.7
Nb-bool	68.7	71.1
Nb-tf	68.6	71.7

Because the SVM-bool classifier is easiest to explain, we used the SVM-bool classifier to rank the word presence or absence features based on their discriminative power toward separating "more popular" and "less popular" messages on "400 messages" in the SVM-bool classifier. Table 4 lists the words that indicate the "more popular" and "less popular" messages. Surprisingly, "contest", "win", "winner", and "free" all appear as top indicators of "less popular" messages. It means that fans are not that interested in winning the contests, probably because of the low chance to win. On the other hand, after checking context of the top indicators of "more popular" messages, we found that fans are more interested in trying new dishes, for example the word "try" occurred in 15 messages, 14 of which are in the "more popular" category.

Table 4. The most indicative words ranked by the SVM-bool classifier

Category	Top indicative words
more popular	try, have, coffee, deal, come, days, october, one, then, like, order, will, who, crunchy, with, million, very, pumpkin, flavors, new, spice, fourthmeal, but, donate, join, early, medium, fall, favorites, served, support
Less popular	the, week, about, http, win, their, I, watch, from, com, last, free, starbucks, first, can, contest, by, yet, memories, tell, amp, visit, live, almost, family, on, fun, our, meal, photo, video

5 Media Type Analysis

Social marketing messages use not only words but also other information types such as links, videos, or photos. In this section we report a statistical analysis to examine the impact of the message types on their popularity. Most of the messages we collected contain a metadata entry called "Media Type" which is either "status", "link", "video", or "photo". A "status" message has only textual descriptions, usually about the business's updates, status, information, etc. A "link", "video", or "photo" message contains a URL, video, or photo respectively. They may or may not contain text. We conjecture that a restaurant chose a media type when posting a message to Facebook in that Facebook is not likely to manually or automatically assign this metadata for each message. We retrieved the media types from all messages and described their basic statistics in Table 5.

We then performed ANOVA test to examine whether there is significant difference between the normalized numbers of "likes" for the four media groups ($N=670$, five messages did not include media type). The homogeneity test shows that the significance is $p<.001$ in Levene's test, indicating that the equal variance hypothesis should be rejected. Therefore we used Games-Howell for post-hoc testing.

The ANOVA test result in Table 6 shows significant difference between the four media groups in terms of the normalized number of "likes" they attracted from fans, $F(3,666)=9.499$, $p<.001$. Because the data violate homogeneity assumption, we used Welch and Brown-Forsythe to further examine the significance of the result. The result is still significant ($p<.001$): $F(3,188.16)$, $p<.001$ and $F(3,561.42)$, $p<.001$ for Welch and Brown-Forsythe tests respectively. On average "status" and "photo" messages received significantly more "likes" than "link" and "video" messages. A possible reason is that many fans chose not to spend the extra effort the click the links or the videos. “Status” messages are even more popular than “photo” messages, although there is no significant difference between them. There is no significant difference between “link” and “video” types either.

Table 5. Basic statistics of media types

Statistics	Link	Status	Video	Photo
#msg	227	190	38	215
Mean	-.22	.19	-.40	.12
SD	.76	1.15	.51	1.06

Table 6. ANOVA test of media type difference

Media type (I)	Media type (J)	Mean difference (I-J)	Sig.
Link	Status	-.41***	.000
Link	Video	.18	.279
Link	Photo	-.34**	.001
Status	Video	.59***	.000
Status	Photo	.074	.908
Video	Photo	-.51***	.000

Note: *p<.05, **p<.01, ***p<.001

6 Conclusions and Limitations

This research examined the characteristics of social marketing messages that affect their popularity. Because popularity of these messages indicates the effectiveness of the corresponding marketing strategies, this research is expected to lay the foundation in order to build computational models to predict the effectiveness of social marketing strategies based on the popularity-related factors discovered in this study.

Using messages posted by a sample of restaurants on Facebook as a case study, we investigated the relationship between the message popularity (measured by the number of “likes”) and the message content and media type. By combining a number of text mining and statistics methods, we found that (1) restaurants share some common marketing strategies, for example they attempt to promote their unique public images, introduce new dishes, as well as run advertisement campaigns like contests; (2) fairly effective prediction models can be built to predict if a message is more or less popular based on the message’s content; (3) some marketing strategies, like promoting new dishes, are more effective while others, such as running contests, are not, probably because of the low chance to win these contests; (4) media type also affect the message popularity: “status” and “photo” messages are more popular than the “link” and “video” ones, probably because of the extra effort to click or play.

This research has a few limitations that we will improve in the future. First, we have not yet controlled the effect of time on the message popularity. Newly-posted messages were penalized to some extent because of their short exposure time. Second, the number of “likes” captures positive feedback only because no “dislikes” button is available. However fans can leave free-text comments which should be able to disclose both positive and negative sentiment, and thus provide more information regarding the effectiveness of social marketing strategies. We are currently analyzing the sentiment of the messages and their comments. Third, the data set is relatively small because Facebook API returns posts of the past 30 days. However, all research methods we employed in this study can be scaled to larger data set. We will continue collecting the posts and comments regularly for the next six months and re-examine if our conclusions still hold on the larger data set. We will also explore relevant approaches to enhance the performance of message popularity prediction [5, 9].

References

1. Blei, D., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. J. Machine Learning Research 3, 993–1022 (2003)
2. Chung, C.K., Pennebaker, J.W.: The Psychological Function of Function Words. In: Fiedler, K. (ed.) Social Communication: Frontiers of Social Psychology, pp. 343–359 (2007)
3. Domingos, P., Pazzani, M.: On the Optimality of the Simple Bayesian Classifier Under Zero-one Loss. Machine Learning 29, 103–130 (1997)
4. Forman, G.: An Extensive Empirical Study of Feature Selection Metrics for Text Categorization. Journal of Machine Learning Research 3, 1289–1305 (2003)
5. Hong, L., Davison, B.D.: Empirical Study of Topic Modeling in Twitter. In: Proceedings of the First Workshop on Social Media Analytics, Washington D.C, July 25 (2010)
6. Joachims, T.: Text categorization with Support Vector Machines: Learning with Many Relevant Features. In: Nédellec, C., Rouveirol, C. (eds.) ECML 1998. LNCS, vol. 1398, pp. 137–142. Springer, Heidelberg (1998)
7. Killian, K.S.: Top 400 Restaurant Chains. Restaurant & Institutions, 28–72 (2009)
8. Mangold, W.G., Faulds, D.J.: Social Media: The New Hybrid Element of the Promotion Mix. Business Horizons 52(4), 357–365 (2009)
9. Szabo, G., Huberman, B.A.: Predicting the Popularity of Online Content. Communications of the ACM 53(8), 80–88 (2010)
10. Xiang, Z., Gretzel, U.: Role of Social Media in Online Travel Information Search. Tourism Management 31(2), 179–188 (2010)
11. Young, L.: Brave New World. Foodservice and Hospitality 42(11), 24–28 (2010)
12. Yu, B.: An evaluation of text classification methods for literary study. Journal of Literary and Linguistic Computing 23(3), 327–343 (2008)