Learning to Predict Subject-Line Opens for Large-Scale Email Marketing

Raju Balakrishnan  
Data Sciences, Groupon Inc.  
Palo Alto CA USA 94306  
Email: raju@groupon.com.

Rajesh Parekh  
Data Sciences, Groupon Inc.  
Palo Alto CA USA 94306  
Email: rajesh@groupon.com.

Abstract—Billions of dollars of services and goods are sold through email marketing. Subject lines have a strong influence on open rates of the emails, as the consumers often open emails based on the subject. Traditionally, the email-subject lines are compiled based on the best assessment of the human editors. We propose a method to help the editors by predicting subject line open rates by learning from past subject lines. The method derives different types of features from subject lines based on keywords, performance of past subject lines and syntax. Furthermore, we evaluate the contribution of individual subject-line keywords to overall open rates based on an iterative method—namely Attribution Scoring—and use this for improved predictions. A random forest based model is trained to combine these features to predict the performance. We use a dataset of more than a hundred thousand different subject lines with many billions of impressions to train and test the method. The proposed method shows significant improvement in prediction accuracy over the baselines for both new as well as already used subject lines.

Keywords-email; deals; subject; learning

I. INTRODUCTION

Emails are an important digital marketing channel behind only search and display ads in marketing dollars spent [8]. In addition to increasing direct marketing through emails, emerging deep discounting markets selling billions of dollars rely on emails to reach out to consumers [11, 1, 10]. Email-based purchasers generally first open the email, then click the links in the email content to purchase. The subject lines of emails have a strong influence on the consumer’s decision to open or not to open the email. As opening the email is a necessary first step for the marketing to be effective, revenue critically depends on the open rates, and consequently on subject line quality. Subject lines are generally written by human editors, based on the content and type of the emails to be sent. Accurately predicting the open rate of a subject line will enable the editors to compile high-performing subjects.

In this paper we propose a method to predict open rates of an email based on the subject line by learning from past subject lines and emails.

Predicting open rates of emails based on the subject lines is generally hard. Depending on how many times a subject line has been used in the past, forecasting its performance may be anywhere from easy to very hard. On the easy side, if an exact subject line has been used in the past in a large number of email campaigns, the open rates of the subject line gives us a statistically significant estimate of how well the subject line will be performing. On the other hand, if the subject line has not been used in the past, predicting the open rates is very hard. In between these two extreme cases, subject lines which are used rarely in the past, and those similar to the previously used subject have varying degrees of difficulty in predicting the open rates.

Apart from sparsity of usage in the past in typical marketing scenarios, another difficulty in predicting open rates is noise due to external factors like location, time of the year, competitor emails, etc. The same subject line—like Starbucks Coffee—may get different open rates in Chicago-Illinois and Scottsdale-Arizona. Many of these external factors will add to the variance in open rates for the same subject lines. These variances impose a maximum upper bound on predicting open rates solely based on subject lines. Further more, predictions should be robust against the noise due to these factors.

Our prediction of open rates depends on three types of features: syntactical, historical, and derived. Syntactical features are directly extracted from the subject lines—like the number of keywords, punctuation marks, etc. Historical features are based on the open rates of the individual keyword in the past. These historical features are derived based on the mean open rates of the keywords with appropriate smoothing and extrapolations. The third group of features—derived features—are also dependent on the historical performance of the subject lines and keywords. But they are derived based on a collective analysis of historical features, instead of analysis of subject lines in isolation. For instance, in a subject line like Starbucks Coffee the influence of the two keywords—Starbucks and Coffee—on the open rates will be different. The high or low open rate for such a multi-word subject is generally caused by one of the keywords, and needs to be attributed to that particular keyword. The derived feature Attribution Scoring tries to infer the contribution of each keyword towards the open rate of the subject line, and assign a score to the keywords accordingly.

To formulate the overall predictions, we train a random forest regression model with these three groups of features as inputs. Many of the interactions between features are modeled as the convolution of the independent features. The final prediction make use of a random forest with six hundred

A US patent is pending on the method described in this paper.
trees and forty one input features. The dataset consists of email corpora containing 107,503 email campaigns with many billions of impressions.\footnote{The exact number of impressions are not given due to confidentiality reasons.} These are United states campaigns in different geographies spanning over several months. We evaluate the predicted open rates against the observed open rates in a test set kept apart from the training set. In addition to the prediction accuracy on the entire test set, we compared the methods against the baseline over separate test sets with varying number of usage-counts in the past to assess the robustness.

\section{Background}

Recently, email campaigns have become increasingly popular as a direct marketing channel for the multi-billion dollar deep-discount market. Typically, emails containing multiple discount offers are sent to consumers in a batch during a particular time of day.

Generally, such a deep-discount email contains number of components as shown in Figure 1:

1) \textbf{Email sender}: e.g. Groupon [10], LivingSocial [11] etc.

2) \textbf{Subject line}: Typically a short description of the email content in a few words, e.g. \textit{Half-off Pizza}. We describe each word in a subject line as a keyword. There may be thousands of emails sent to different subscribers with the same subject line. Each email with a given subject line is an impressions of the subject line.

3) \textbf{Email content}: Generally contains number of deal links with pictures and descriptions of deals shown in the open view in Figure 1. The consumer may click on these deal links to get more details about the the deals and to make purchases.

Two important pieces of information the consumers have about the email at first glance—i.e. before opening as in the inbox view in Figure 1—are the sender, and the subject line. Therefore, a consumer’s decision to open the email critically depends on these two factors. Since opening the email is a necessary precondition for purchase, improving open rates is critical to business success. For our dataset and the use case, the sender name is uniform across all the emails. Hence we assume that the only variable aspect influencing a consumer’s decision to open the email is the subject line.

To give an overview of the process by which email subjects are compiled, the subject lines are written by human editors. Although the emails may contain multiple deals, the subjects are based on the deal at the top position in the email (deal at the top left for two column formats). The editors may write the email subjects based on the deal title, deal description, past email subjects, etc. These subjects are used for emails for which the deal appears in the top position.

The same type of deal may receive different subject lines. Subject lines are indicative of deal content but may have variations within this condition. For example, one pizza deal may be sent with the subject line \textit{Up to Half off Pizza} while another pizza deal (from a different merchant) may be sent with a subject line \textit{One Topping Pizza}. A subjects reflective of the email content is necessary for long-term consumer satisfaction and retention.

\section{Predicting Open Rates}

We model the subject lines as an ordered set of keywords. The features are derived at a keyword level, and the order of the keywords is taken into consideration for prediction. In this model, the problem may be stated as to predict the open-rate $o_i$ of emails with the subject line $s_i$ where, $s_i = \langle k_1, k_2, \ldots, k_n \rangle$ and $k_j$ is the $j^{th}$ keyword in $s_i$ and $n$ is the total number of keywords. For predicting the open rates, we derive a number of features for each keyword in the subject line, and the subject line as a whole. The twenty-one stand-alone features are listed in Table I. The features fall into three broad categories—syntactical, historical and derived—as described in the following sections.

\subsection{Syntactical Features}

These features are purely syntactical, and straightforward to derive. These features try to capture consumer’s preference for syntactical attributes—for example short vs. long subjects.

\subsection{Historical Features}

These features are derived based on the past performance of the keywords in the subject. The set of features are based on the hypothesis that past performance of the keywords is indicative of open rates in the future.

\textbf{Mean open rates of keywords in the subject}: We compute the open rate of a keyword as the open rate of all the subject

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig1.png}
\caption{Sample daily deals mail view in the inbox and after opening the mail. The sender and the email subject are the two prominent fields the user would see in his inbox.}
\end{figure}
lines containing keywords. We make use of up to first four keywords from the left in the subject line. For the subject lines shorter than this, we impute the open rates with the the mean open rate of all the keywords. For longer subject lines—which are very rare in our corpus—the other keywords are ignored. We smoothed the open rates of keywords to account for difference in number of subject lines containing the keyword. Accordingly OPR of the $i^{th}$ keyword $k_i$ is computed as,

$$ \text{opr}_{i} = \frac{1 + \text{opr}_{j} N}{N + 1} \quad (1) $$

where $N$ is the number of email campaigns in which the $k_i$ has occurred.

For new keywords, i.e. keywords which are not used in the past, we filled the open rate as the weighted-average open rates of the most-similar keywords. Similarity is computed using the Levenshtein distance (edit distance). If more than one keyword is found in the equal minimum distance, we take the weighted mean of the open rates of these keywords, where weight equal to the number of subject lines in which the keyword was used.

**Occurrence score:** The mean open rates of the keywords described above indicates the overall performance, but signals indicating consistency in performance of a keyword are lost in the process of averaging. For example, the mean open rate of a keyword that performed well only in a few campaigns will be almost the same as the mean open rate of a keyword that performed equally well in all the ten campaigns. Though the smoothing may capture some of this difference, it is desirable to include a stronger signal indicating consistency. Occurrence score captures the signals indicating the consistency in performance of the keywords.

To derive the occurrence scores, we divide the campaigns into two sets. The first set contains all the email campaigns preforming better than the median open rates. The second set consists of all the campaigns doing worse than the median open rates. The occurrence score of a keyword is computed as the difference between the number of times the keyword occurred in the first set and the second set. For example, if $Pizza$ occurred eight times in campaigns preforming better than the median open rate and four times in campaigns preforming worse than the median open rate, the occurrence score for $Pizza$ is four. Similar to the historical open rates, we used the mean occurrence score of the nearest keywords for new keywords. Instead of taking the weighted mean, we computed the simple mean of the nearest keywords for the occurrence score. We do not use the weighted mean, as the number of subject lines in which a keyword was used does not directly account for the certainty of the occurrence score.

**Count score:** The number of subjects in which the first keyword in the subject appeared in the past. This is based on the hypothesis that editors are likely to use the high performing keywords more frequently, hence the keyword popularity may indicate performance. We used count score of only the first keyword. Although we tried using count score for all the four keywords, the other keyword positions were found to have no effect on accuracies.

**C. Derived Features: Attribution Scoring**

Attribution scoring tries to infer the contribution of the individual keywords in the subject line to the open rate of the email. For example, if an email with a subject line *Half off Starbucks Coffee* has high open rates, it is possible that most people who opened the email are attracted by one of the keywords, and the other keywords did not have much influence. While a simpler metric like the mean open rate ignores this difference and assigns the high open rates equally to all the keywords, attribution scoring tries to infer the influence of different keywords by a collective analysis of subject lines.

An initial idea for inferring the influence of different keywords in a given subject line is to look at other subject lines

<table>
<thead>
<tr>
<th>Group</th>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Syntactical</strong></td>
<td>Number of keywords</td>
<td>Number of space separated keywords.</td>
</tr>
<tr>
<td></td>
<td>Number of characters</td>
<td>Length in characters.</td>
</tr>
<tr>
<td></td>
<td>Number of short keywords</td>
<td>Keywords in length less than four characters.</td>
</tr>
<tr>
<td></td>
<td>Has digits</td>
<td>Whether subject line contains digits (0-9).</td>
</tr>
<tr>
<td></td>
<td>Has punctuation</td>
<td>Whether subject line contains punctuation marks (e.g., ., !).</td>
</tr>
<tr>
<td><strong>Historical</strong></td>
<td>Open Rate: first keyword</td>
<td>Open rate of the first keyword from the left.</td>
</tr>
<tr>
<td></td>
<td>Open rate: second keyword</td>
<td>Open rate of the second keyword.</td>
</tr>
<tr>
<td></td>
<td>Open rate: third keyword</td>
<td>Open rate of the third keyword.</td>
</tr>
<tr>
<td></td>
<td>Open rate: fourth keyword</td>
<td>Open rate of the fourth keyword.</td>
</tr>
<tr>
<td></td>
<td>Occurrence score: first keyword</td>
<td>Occurrence score of the first keyword from the left.</td>
</tr>
<tr>
<td></td>
<td>Occurrence score: second keyword</td>
<td>Occurrence score of the second keyword.</td>
</tr>
<tr>
<td></td>
<td>Occurrence score: third keyword</td>
<td>Occurrence score of the third keyword.</td>
</tr>
<tr>
<td></td>
<td>Occurrence score: fourth keyword</td>
<td>Occurrence score of the fourth keyword.</td>
</tr>
<tr>
<td></td>
<td>Count score: first keyword</td>
<td>Number of times the first keyword was used in the past.</td>
</tr>
<tr>
<td><strong>Derived</strong></td>
<td>Attribution score: first keyword</td>
<td>Attribution score of the first keyword from the left.</td>
</tr>
<tr>
<td></td>
<td>Attribution score: second keyword</td>
<td>Attribution score of the second keyword.</td>
</tr>
<tr>
<td></td>
<td>Attribution score: third keyword</td>
<td>Attribution score of the third keyword.</td>
</tr>
<tr>
<td></td>
<td>Attribution score: fourth keyword</td>
<td>Attribution score of the fourth keyword.</td>
</tr>
</tbody>
</table>

**TABLE I**

Stand-alone features for the predictor.
to see how different keywords are performing and adjust our belief accordingly. For example, as we mentioned above, in the subject line *Starbucks Coffee*, initially we believe that all the keywords in a subject line contributed equally to the open rate of a subject line. Based on this, we set the score of both the keywords equal to the open rate of the subject line. Subsequently, we check the open rates of keywords in other subject lines. If we observe that many other subject lines with the keyword *Coffee* do not have high open rates whereas subjects with *Starbucks* are generally doing well, our belief that *Starbucks* caused the higher open rate increases. Consequently we reduce the score of coffee by a small amount and increase the score of *Starbucks* by a small amount in this subject line. Notice that this is a two-way process, i.e., when we adjust the scores of keywords in the $i^{th}$ email $e_i$ based on scores of keywords in $e_j$, we adjust the scores of keywords in $e_j$ based on the scores of keywords in $e_i$ as well. Hence we have to iterate over this score adjustment multiple times to get the final attribution scores.

To compute the attribution score, we start by building an email-keyword voting matrix where rows are the emails and columns are the keywords. If the $j^{th}$ keyword $k_j$ occurs in the $i^{th}$ subject line $s_i$, element $o_{ij}$ of the voting matrix is set to the open rate of the email. If the keyword $k_j$ is not occurring in $s_i$, $o_{ij}$ is set to the mean open rate of all the keywords. Rates, it may be the case that one of the keyword drives the higher open rates, while the other keyword is inconsequential. Attribution scoring try to guess this factor based on an iterative matrix multiplication process.

We implement the attribution scoring as an iterative matrix multiplication. Let $V^i$ denotes the voting matrix after the $i^{th}$ iteration. Let $P^i$ denotes the scores vector of the keywords, where $|P^i|$ is the total number of keywords. The computation is as following,

$$S^{i+1} = V^i P^i$$  \hspace{1cm} (2)

In the second step the votes of the emails are modified as,

$$V^{i+1,*T} = V^{i,*T} P^{iT}$$  \hspace{1cm} (3)

where $V^{i,*}$ is the $i^{th}$ row of $V^i$. We repeat this process till the scores converge. Like other scores, for new keywords we used the mean attribution scores of the nearest neighbors.

d. Combining Features

We use a random forest regression to combine the scores from different sources to compute the final score. The features described above were scaled to numbers between zero and one by subtracting the minimum value followed by division by the range of the feature values. In addition to features listed in Table I, we explicitly modeled the interaction between many features as convolution of feature vectors. Thus we used a total of forty-one subject line features. We used a total of six hundred trees in the random forest, and used the R implementation in randomForest package. Preprocessing and feature engineering are implemented in Java, SQL on Teradata [15] and R. In addition to random forest, we tried other regression models like support vector regression, linear regression, and logistic regression. The random forest performed the best compared to other models.

IV. EXPERIMENTS AND RESULTS

We evaluated the proposed method on a large email dataset with billions of email impressions. The proposed method is compared with baseline on the ability to predict open rates of new (subject lines never used in the past) and used (subject lines used one or many times in the past) subjects. Our evaluations estimate multiple measures of the effectiveness of each method. The proposed method consistently performs better than the baseline for both the new as well as the used subject lines.

A. Dataset

The email dataset consists of 107,503 different subject lines of daily deal emails. The dataset consists of email impressions spanning a few months after removing subject lines having less than 1024 impressions. As each subject line gets several thousands of impressions, these subjects get a total of several billions of impressions. These are email campaigns in North America, and contain a mix of geographical areas covering a major portion of the country including rural and urban areas. The number of emails sent and number of opens for each subject line and email are recorded in the dataset.

Considering the content of the email, an email may contain multiple daily deals as shown in Figure 1, most of them pertaining to local merchants. The subject lines of the emails are derived from the top deal in the email. Generally the

<table>
<thead>
<tr>
<th>Unweighted Measures</th>
<th>Random Forest</th>
<th>Lookup Table</th>
<th>Improvement (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Root mean square error</td>
<td>$1.60 \times 10^{-2}$</td>
<td>$1.78 \times 10^{-2}$</td>
<td>10.11%</td>
</tr>
<tr>
<td>KL-Divergence</td>
<td>5.40 $\times 10^{-3}$</td>
<td>7.24 $\times 10^{-3}$</td>
<td>25.41%</td>
</tr>
<tr>
<td>Correlation</td>
<td>0.352</td>
<td>0.315</td>
<td>11.74%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Measures Weighted by Impressions</th>
<th>Random Forest</th>
<th>Lookup Table</th>
<th>Improvement (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weighted MAE</td>
<td>$1.4 \times 10^{-2}$</td>
<td>$1.38 \times 10^{-2}$</td>
<td>17.39%</td>
</tr>
<tr>
<td>Weighted Root Mean Squared Error</td>
<td>$1.56 \times 10^{-2}$</td>
<td>$1.86 \times 10^{-2}$</td>
<td>16.13%</td>
</tr>
</tbody>
</table>

**TABLE II**

Comparison of the proposed measure and the baselines. The unweighted measures gives equal importance to all the subject lines whereas weighted measures weigh the subject lines by the number of impressions. The proposed predictions performs better than the baselines in all the measures.
subject lines do not contain merchant names, but only a
description of the deals. The audience for each emails is
primarily selected based on the location of the merchant,
consumer, and type of the deal, etc. From this dataset we
randomly selected a set of 10,750 email subjects as the test
set, and rest of the data set was used as the training set. All
the results shown below are on this test set.

B. Baselines

To the best of our knowledge there is no existing work
on predicting open rates of emails based on subject lines.
Hence we used an intuitive lookup table based predictor as
the baseline to compare our method. The open rate of a
subject line is predicted equal to be the mean open rates
of the same subject line in the past. If the subject lines has
not been used in the past, the open rate is predicted to be
equal to the average open rate of all the subject lines. We
compare the methods in different sets of emails including rare,
popular as well as new subject lines. The proposed method
performs better even for subject lines frequently used in the
past, we compare accuracies of other regression models with
the proposed random forest method.

C. Results and Analysis

The first set of experiments compares the proposed method
to the baseline quantitatively. We compute root mean squared
errors, KL-divergence, and the correlation between the pre-
pdicted and the observed open rates, and the results are listed
in Table II. The weighted measures weight the prediction
accuracy of subject lines by the total number of impressions,
whereas the un-weighted consider all the subject lines equally.
The weighted measures are important as the business impact of
accuracy of prediction of a given subject line increases with the
number of impressions. In weighted measures, we compute the
weighted RMS error and the weighted Mean Absolute Error
(MAE) for the test set. The proposed method performs better
than the baseline in all the measures. For all the measures
in Table II, the proposed measure exceeds the baseline with
significance level of 0.95 or more. Note that the increase in
accuracies is as high as 25% by some measures. Regarding the
seemingly low correlation values, please keep in mind that the
open rates is sensitive to many external factors not included as
an input variables—not just the subject line—as we described
in Section II. These external factors imposes an upper limit
on accuracies on any predictions solely based on the subject.

In a second experiment, we evaluate the prediction accuracy
on rare and new subjects. For these evaluations we selected
two subsets from the test set. New subjects are a set of subjects
never used in the past, i.e a set of subjects that did not appear
in the training set. 44.2% of the subject lines falls into this
category. We use fewer buckets in Figure 2 as the total number
of subject lines is smaller. Rare subjects are the ones used less
than five times in the past. 74% of the email subjects falls in
this category. For both categories, the mean predictions are
monotonically increasing with the mean open rates, showing
the effectiveness of the model. These two evaluations indicate
the generalization capability of the model. Note that the basic
lookup table cannot be used for the new subject lines at all,
hence cannot be compared with the proposed method.

V. RELATED WORK

A related area of research is in features of search result
snippets in web search engines leading to higher CTRs. Yue
et al. [16] studied the effect of bolded query keywords in
the snippets using Fairpairs [13] algorithm to cancel for the
position bias. Clarke et al.[6] studied influence of caption
features of web search results and showed that the simple
features like the length and number of matching query terms
affects the click-through rate. The influence of visual features
in image search has been studied by Geng et al. [9], showing
that the image attractiveness has a significant influence on
user satisfaction and is helpful in improving index coverage.
Though these studies are indicative of features attracting
people’s attention, email subjects are significantly different
from search results, as there is no search involved, and as the
attention pattern in an email inbox is significantly different
from that over a search result page [4].

There has been considerable research and analysis in the
area of daily deals. Byers et al. [5] analyzed the effect of
running daily deals on reputation and rating in web and
social media. With the increasing popularity of daily deals,
optimal bidding and scheduling strategies for daily deals
type guaranteed ads have been analyzed [3]. The effect of
daily deals on business profitability and popularity has been
analyzed [7]. Although dealing with different aspects of daily
deal business, none of these papers consider the problem of
improving the email subjects for daily deals we are considering
in this paper.
The application of email as a marketing medium, as well as as a medium for customer relations has been identified and exploited even in early years [14]. In viral marketing, the aspects increasing the probability of emails to be forwarded has been studied [12]. But unlike our method of predicting open rates from directly derived syntactic and semantic features, these viral marketing studies focused on abstract high-level features like ability to evoke emotions, frequency of forwarding by senders etc. The effect of the visual appearance of display ads on click-through rates and conversions has been studied [2]. The independent effect of targeting and obtrusive visual design along with the mutual influence of these two factors were studied by Goldfarb et al. The study concludes that visual obtrusiveness may reduce the click-through rate for contextual ads directly collected by surveys. Although there is a considerable amount of related research in the area of search, display ads, daily deals, and viral marketing as discussed above, there is no work focusing on predicting email-subject quality addressed in this paper to the best of our knowledge.

VI. CONCLUSIONS AND FUTURE WORK

Compiling relevant subject lines appealing to consumers is important in multi-billion email marketing. Predicting the open rates of a subject will help to compile better subject lines leading to higher email open rates. We propose and implement a prediction method based on syntactical, historical, and derived features of the email subjects. The syntactical features include length of the subject lines, punctuation, and number of keywords. The historical features include the past open rates of different keywords in the subject line. The derived features try to infer the contribution of individual keywords to the open rates of the subjects in the past based on the collective analysis of the email subject-lines. Specifically, we propose an iterative method to assess the contribution of individual keywords to the open rates of emails—namely Attribution Scoring. Our model is trained and tested on a dataset containing billions of email impressions. The proposed model shows significant improvement over the baseline in prediction accuracies both for new subject lines as well as subject lines used a number of times in the past.

Considering related problems, personalizing email subject lines is a promising direction to improve attractiveness to the user. Instead of sending the same email subject to every consumer as we describe here, personalized subject lines may be derived based on what aspect of the email content is likely to best appeal the specific user. With respect to extensions to the related domains, improving click-through rate and attractiveness is a common problem in in compiling better captions and titles for online videos, display ads, search results etc. The methods described here may be extended and applied for these related data types.

REFERENCES