Tweetalyst: Using Twitter Data to Analyze Consumer Decision Process

Viraj Kulkarni, Suryaveer Singh Lodha, Yin-chia Yeh

Abstract

Marketers are increasingly turning to social media platforms to extract insights into consumer behavior. Tweetalyst is an attempt to use microblogging platforms such as twitter to identify users at different stages of the decision process of buying a given product. In this report, we introduce related concepts and describe the approach and working of Tweetalyst. We describe the datasets we use and present our results. We conclude by discussing our experiences and the insights we gained by working on this project.

Introduction

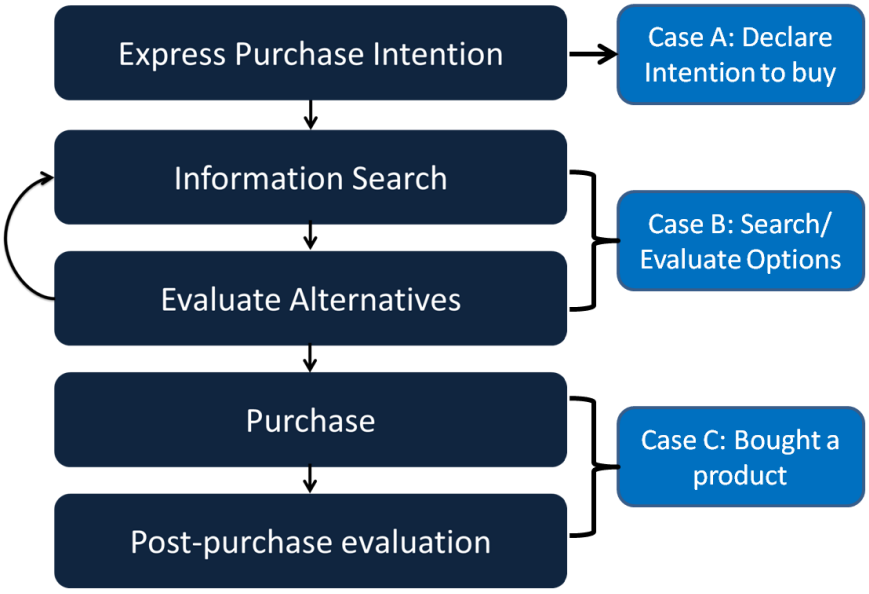
The number of users who microblog their daily activities is increasing rapidly. Twitter provides unique opportunities for behavioral data analysis since it has a large base of active users who tweet often and the behavioral data that is generated is made publicly accessible by Twitter. The large amount of data available makes it possible to use data mining techniques to gain insights into consumer behavior.

Consumer decision processes are the decision making processes undertaken by consumers in regard to a potential market transaction before, during, and after the purchase of a product or service. A number of models exist for these processes and they have been studied for some time in the fields of economics, psychology and marketing. There are economic models that are based on quantitative approaches and the assumption of rationality in buyer behavior. Psychological models generally qualitative and concentrate on psychological and cognitive processes such as motivation and need recognition.

A number of these models have been studied extensively in the context of the offline world. Our project aims to validate their existence in the online world and study the variations in the offline and online buying processes. For example, in the offline world, the consumer would typically go to the store to buy a product and have direct interaction with the salesperson about his needs from the product and budget, and evaluate

options with the salesperson. We are trying to see how does a model like this work in the online world, where consumers increasingly buy products online.

We study the offline model in the adjoining figure. This model describes five discrete stages in a typical consumer buying process. After studying this cycle in detail and based on our pilot studies on Twitter, we decided to consider 3 discrete steps in online buying behavior. In offline world, when a consumer expresses intent to buy or recognizes need, it similar to the case on Twitter when someone tweets about willing to buy/ intending to buy a service/product. The next steps in offline world are Information search and evaluation of alternatives. Here, the consumer searches for more information about the product/service, and in this process, may uncover alternatives. We observe that it is often a back and forth process and at least on Twitter it is difficult to tell these two stages apart even for humans. Hence, we decided to combine these two offline steps into one step online – Case B. Finally, when the consumer does perform a purchase, it is relatively easy to track in offline world. However, post purchase evaluation is rare in offline world, and mostly only bad experiences are reported often in offline world. For online buying trends, we find it difficult to clearly distinguish between these steps and hence combine both of them together to develop a Case C. Case C is a collection of all tweets made after a purchase.



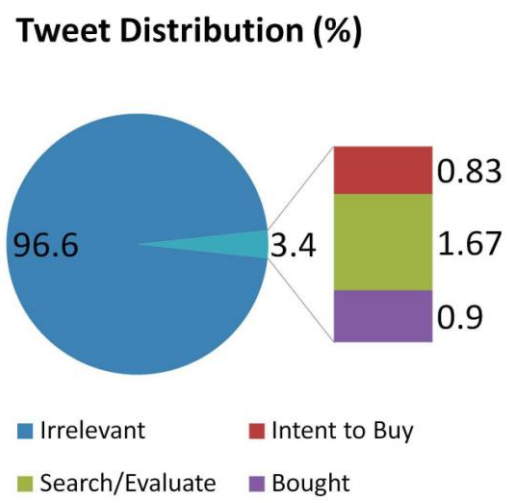
Dataset Collection

As we did not have any training set of twitter data, we needed to create annotated data ourselves. We needed two typed of annotation per tweet – does it represent a consumer buying behavior pattern (Yes/No – Binary classification). If a tweet does represent a consumer buying pattern, we try to assign it to a particular class - Class A, B or C, as discussed above.

Pilot test

Initially we collected 10,000 tweets (using both MarkLogic and the search API, using scala). These 10,000 tweets were a random set of tweets and we made sure that all

these tweets were in English. The language information is embedded in the tweet itself ('iso_language_code' == 'en' and 'lang' == 'en'). We also created a hash map of all the tweets with key as the tweet_id, so that we only look at unique tweets in English language. Then our group members manually classified each tweet as belonging to consumer buying behavior or not. We first decided to do a coarse classification (and not the multivariate classification) as we wanted to validate our hypothesis that enough consumers do tweet about their thoughts when buying products. Also, we wanted to learn what kind of products people tweet about before/after buying. In that respect, we think that this manual task of classification was very helpful, as it not only gave a good idea about the kind of data we will be dealing with, but also gave us insights into the usage of twitter by potential consumers and helped us make the next step of data collection highly efficient. Based on our pilot tests of these 10,000 tweets, we found that 3.4% of the tweets were related to consumer buying behavior. We further classified the relevant tweets into the three classes, as shown in the adjoining figure.



Crowdsourcing approach

We also tried to use crowdsourcing techniques to help us annotate and collect data. As we felt that data annotation is very crucial to our models, especially as we would have limited data (in the order of 10,000 or so tweets), we decided against using crowdsourcing for data labeling. One of the key roadblocks in implementing crowdsourcing approaches is to develop mechanisms to validate a worker's response. As we didn't have enough time and resources to develop such a mechanism, we decided to utilize crowdsourcing for collecting tweets. We posted tasks on Amazon Mechanical Turk where workers were asked to submit 7 tweets from their twitter graph which belonged to either of these consuming buying behaviors – declaration of intent to buy, search or evaluation of alternatives, and declaration of a recent purchase. For each HIT, we provided workers with a few examples of tweets which belonged to that category. For each approved we paid the workers 0.05$. We periodically posted 300 tasks every hour for almost a week, and then manually read worker submissions and approved helpful/ meaningful submissions. Even though we did get some good responses, a majority of the responses were not helpful. We also found some extreme instances where the crowd worker made 7 posts on twitter while completing the HIT. While this did give us some useful search keywords, this is one check we might want to implement in a future system which automatically validates worker's work for such HITs.

Final Dataset creation

Based on our initial tests on twitter data during the pilot study and crowdsourcing approach, we found that there were certain product categories which were tweeted about most. We also found good key search terms for finding buying behavior related tweets and learnt about the type of tweets to ignore while collecting more. We found that most of the relevant tweets were about – cars, tablet laptops, ipad, smartphones and cameras. From the manual analysis we also found that certain verbs - bought, purchased, thinking of, cheaper were quite common. Now, because we wanted to gather enough tweets so as to have enough tweets for each class of buying behavior, we decided to perform more specific searches on the twitter dataset. Hence, we first identified a few product categories along with specific search keywords for such categories. We show a few of them below:

Example of products tried:

(a) Smartphones - smartphone, iphone, android, windows phone, samsung phone, htc phone, galaxy phone, blackberry

(b) Cameras - dslr, camera, point shoot, nikon, canon

(c) Cars - car, honda, toyota, civic, camry, accord, prius, nissan, sedan, sports car, impala, chevrolet, mazda, ford, bmw, mercedes

(d) Tablets - tablet, laptop, ipad, touchpad, notebook, dell, macbook, sony vaio, hp

Example of a set of verbs - bought, purchased, thinking of, cheaper, pay, looking for, thinking about, want to buy, recommend, suggest, advice, think of, budget, discount, deal, bargain.

We define a query Q on Twitter search API as <verbTerm> <productTerm>. An example of such a query: “looking for iphone”. We run our program through all the combinations of verbTerms and productTerm. As discussed above, we maintain a dynamic map of tweets, with tweet_id as the key, so that we never get any duplicate tweets. We only search only for “English” tweets. Also, while searching for keywords, we first convert the entire tweet into lower case, as we learnt that the Twitter Search API is case sensitive and affected our regex (Regular Expression) based search. Based on these modified (biased) search terms, we collected 6000 more tweets and manually labeled the data. We found that approximately 43% of these were relevant and ended up with 2600 relevant tweets in total.

Approach and Results

Given the manually labeled dataset, we develop a classifier that can detect tweets relevant to consumer buying behavior. Since our pilot study shows that certain tweets usually contain keywords such as “bought”, “suggestion”, or “think of,” we think a naive Bayes classifier might be a good fit to catch the characteristics of these tweets.

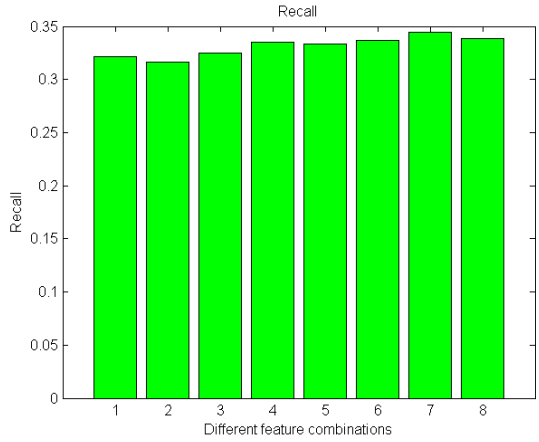
We train a naïve Bayes classifier with multinomial model and Laplace smoothing. We also try a combination a different features, including stop word removal, stemming, and unigram/bigram. The fixed parameter is multinomial model and Laplace smoothing. The results are shown below.

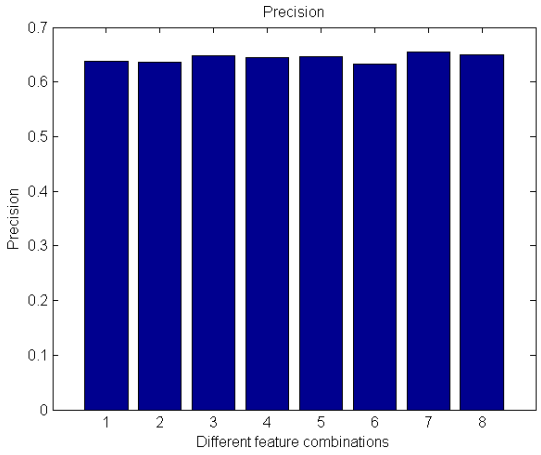
Top Unigram Terms (with Stopword removal and Porter Stemming)

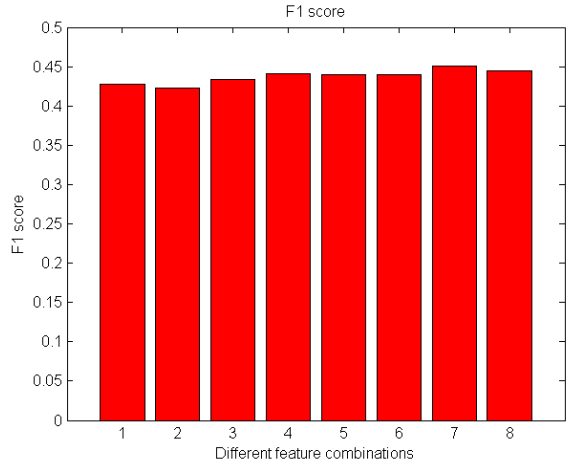
Positive	Weight	Negative	Weight
phone	0.00039	http	0.00260
android	0.00033	Microsoft	0.00117
want	0.00032	Google	0.00100
iphone	0.00030	RT	0.00088
buy	0.00028	apple	0.00081
think	0.00027	Android	0.00053
purchase	0.00024	Nexus	0.00028
recommend	0.00021	Samsung	0.00025
suggest	0.00015	Cream	0.00022
new	0.00012	Ice	0.00021

Top Unigram Terms (with Stopword removal and Porter Stemming)

Positive	Weight	Negative	Weight
want to	7.907E-5	http t	0.00160
to buy	7.755E-5	Ice Cream	0.00013
i think	4.410E-5	Galaxy Nexus	0.00012
i want	4.258E-5	Cream Sandwich	0.00012
do you	1.976E-5	I think	8.392E-5
for android	1.976E-5	to buy	6.714E-5
to purchase	1.976E-5	Google nexusprime	4.906E-5
iphone s	1.976E-5	Android Google	5.551E-5
i recommend	1.824E-5	Steve Ballmer	4.454E-5







Legend
1: Unigram/No Stemming/No stop word removal
2: Unigram/No Stemming/Stop word removal
3: Unigram/Stemming/No stop word removal
4: Unigram/Stemming/Stop word removal
5: Bigram/No Stemming/No stop word removal
6: Bigram/No Stemming/Stop word removal
7: Bigram/Stemming/No stop word removal
8: Bigram/Stemming/Stop word removal

We see that bigram with stopword removal and Porter Stemming gives us best results. It is also interesting to note that some of the top unigram negative words also appear in the top negative bigram words. Also, unigrams such as “ice” and “cream” do not make much sense by themselves, but in the bigram model – ‘ice cream” and “cream sandwich” help us in understanding that those are terms related to smartphone. We also see that most of the RT (re-tweets) and http (web addresses) contribute mostly to the negative class in our Naïve Bayes classifier. Presence of “android google” and “steve ballmer” in negative set also suggests that there is probably a lot more twitter based marketing of Microsoft and Android based smartphones, as compared to iPhones. It was good to see search terms such as “want to”, “I want” etc in positive category. Our best Precision score is 66.24% and Recall score is 34.69%. We performed a 10-cross validation on our dataset. We do not get high Precision and Recall scores. We believe that is due to the fact that our training database is highly biased and opinionated, as discussed in the next section. We also get 52% L2 regression scores for multivariate classification, which again are not high/ promising, but it would be interesting to check run these algorithms on larger datasets.

Discussion

Based on our endeavor to develop a decent training data set and application of a few standard classification techniques such as Naïve Bayes and L2 regression, we did learn quite a few things which are peculiar to tweet data. As we also tried similar approaches on larger data sets (which were pre annotated for us, unlike this twitter data), it was insightful to recognize aspects of twitter data which are both similar and different from IMDB review database and Amazon reviews database.

We find that the amount of advertisements on twitter is quite a lot. As we are trying to identify consumer buying behavior related tweets, it is critical for us to be able to distinguish between tweets posted by consumers as opposed to marketers. This problem is complicated by the fact that most of the top keywords such as “budget”, “buy” etc are common to tweets posted by both consumers and marketers. Based on our model, we learnt that keywords such as “budget”, “http” etc as unigrams, “iphone android htc samsung” as n-grams is good indicator for tweets posted by marketers. Also, we find that if a tweet is retweet-ed (RT) a bunch of times in succession, it mostly relates to a marketer post, or a trending blog/review topic. It was hard for us humans as well while labeling if this tweet should be considered as a consumer buying behavior related tweet or not. We think, one approach which might work well would be to make a list of user_id of marketers on Twitter and avoid tweets from those user ids.

We also found manual labeling of twitter data as Class A/B/C to be biased and opinionated. After having classified close to 17000 tweets in a group 3 people, we believe that at times it is difficult for one person to accurately classify a tweet in a separate class. Initially we started with 5 classes and then resorted to 3 major classes because it was difficult for us to differentiate ourselves if a tweet was about asking for

suggestions or evaluating an option. Often, twitter users include both sentiments in a single tweet, for e.g. *“I’m thinking to buy a Canon T2i, but haven’t researched much. Can a Nokia/Canon pro help me out”.* We think, it would have been better if more than 3 people had worked on each tweet and classified it into categories, so as to have a more robust and useful training dataset. We could also try crowdsourcing techniques for such tasks, but we didn’t venture deep into that because of limited resources.

One other direction of future work is to analyze consumer’s tweeting behavior with respect to different products. It would be interesting to see if people tweet differently for different product. For example, do people ask for more recommendations when buying a car than buying a camera?