Understanding the Functions of Business Accounts on Twitter

Ana-Maria Popescu
Yahoo! Labs
Sunnyvale, CA, 94089
amp@yahoo-inc.com

Alpa Jain
Yahoo! Labs
Sunnyvale, CA, 94089
alpa@yahoo-inc.com

ABSTRACT

This paper performs an initial exploration of business Twitter accounts in order to start understanding how businesses interact with their users and viceversa. We provide an analysis of *business tweet* types and topics and show that specific business tweet classes such as *deals* and *events* can be reliably identified for customer use.

Categories and Subject Descriptors

H.4.0 [**Information Systems**]: Information Systems Applications—*general*

General Terms

Algorithms

1. INTRODUCTION

Businesses are increasingly viewing social media and microblogging services as a platform for reaching out to their customers as well as seeking out and adding new customers. In this paper, we perform an initial exploration of business Twitter accounts in order to understand how businesses interact with users. We believe this would help differentiate successful business-related tweets from unsuccessful ones, offer effective customer facing strategies to businesses, or recommend business-related content to Twitter users. We make the following contributions:

1. We provide a first analysis of business tweet classes derived by manually analyzing 1000 business tweets from 5,245 business accounts.

2. We report encouraging results on the tasks of identifying tweets from the *deal* and *event* classes.

3. By analyzing *topics* from 3 months' worth of business tweets, we find that good quality, easily recognizable topics of general interest can be retrieved for potential use in organizing business tweets.

2. CLASSES OF BUSINESS TWEETS

Our first step is to build a taxonomy of business-related tweets. We start by identifying a set of business accounts with the help of a dictionary of business names extracted from Wikipedia and Y! Local, together with the Twitter user data from a Twitter firehose from Jan. 2010 through June 2010. We mark as business accounts those *verified* Twitter acounts whose name and URL match the dictionary entry (e.g., @*Starbucks*, *HP_PC*, *Oriflame*, etc.), a step resulting in a set of 5,245 accounts. We then analyze a random sample of 1,000 tweets and manually construct a taxonomy of business tweets (see Figure 1 and Table 1 for details).

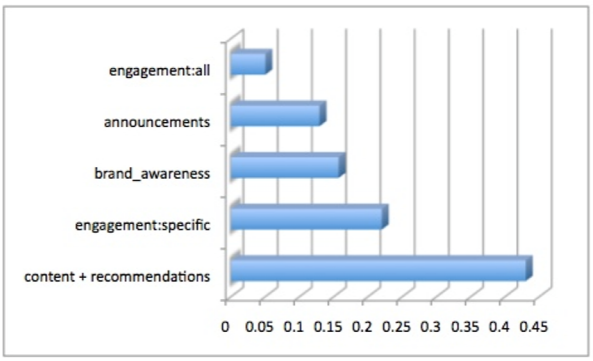

Figure 1: General classes of business account tweets

The largest subset of business tweets is represented by *generic content* typically in the form of *news updates* (many times including links) which are usually about the business domain. More rarely, *opinions* are expressed or *advice* is offered. The second largest subset covers *engagement* with *specific users*, which usually takes the form of *customer service* conversations, although more general discussions are also conducted; businesses also retweet feedback or appreciation messages from specific customers. They also explicitly address their entire audience (*engagement:all*) when conducting surveys or enticing them to participate in contests, usually geared towards gathering more followers. A particularly interesting set of tweets is represented by *announcements*: events, deals or offers, job openings, schedule changes can all be announced via Twitter.

3. DEAL AND EVENT IDENTIFICATION

Our ultimate goal is to automatically monitor and classify business tweets into the categories we previously discussed. In this paper we focus on 2 important classes, *deals* (coupons, offers) and *events*.

Supervised ML We treat the identification of tweets from a given class as a supervised classification problem and employ the Gradient Boosted Decision Trees ML framework to solve it. A tweet is described by a collection of lexical and semantic features. An important resource for feature

WWW 2011, March 28–April 1, 2011, Hyderabad, India.
ACM 978-1-4503-0637-9/11/03.

Engagement: all (0.04)	Examples
survey/poll (0.02)	Is an Alicia Keys written song good enough? decide: http://ow.ly/juuC""
contest (0.01)	follow @miladyteam to win an ipod touch!
other (0.01)	Have a great 2010 !
Announcements (0.13)	**Examples**
events(0.07)	November 21st enjoy a Gingerbread House Workshop in Shakopee MN
deal/coupon/sale(0.03)	Deal of the Day: HP Photosmart C8180 All-in-One Printer / Scanner / Copier 199.99
product release (0.001)	New doubleTwist Windows release is out!..
schedule (0.001)	Closed Monday
other (0.088)	For charities: payment schedule over Christmas
job openings (0.01)	diesel mechanic (New Brunswick)
Brand awareness (0.157)	Best solutions for your scrapbooking needs!
Engagement: specific (0.22)	**Examples**
customer service (0.192)	@bhkahului Do you still have concerns ?
general discussion (0.02)	@*milkyBarKed* I'm a Pet Shop Boys fan
showcase/retweet (0.01)	RT @ralphryan: Personal responsibility Don't blame someone else (...)
other (0.01)	@popjustice what
Content +links (0.43)	**Examples**
news updates (0.37)	Commercial Vehicle Group Reports Second Quarter 2009 Results http://bit.ly/1tLydr
opinions/advice (0.006)	what an incredible goal for team USA woot
other 0.06	what is this

Table 1: Classes and subclasses of business tweets

construction is a class-specific lexicon, consisting of tokens (unigrams, bigrams) correlated with the class of the tweet.

Class-specific lexicon construction: Given a tweet class C and a set S of example tweets of type C, we proceed as follows:

(a) an initial lexicon L is derived by taking the top most k unigrams and bigrams from S;

(b) a sample of 100000 tweets from the 5,245 known business accounts is ranked by a simple Noisy-Or scoring function: $1 - \prod_{i=1}^{|L|}(1 - score(t_i))$, where $t_i \in L$ and $score(t_i)$ is the % of tweet tokens which appear in L.

(c) M = the top m% of ranked tweets is retained (m is estimated using set of tweets in section 1). A topic model for M is built by running a LDA package (Latent Dirichlet Annotation) [1]: all topics T which contain one of a few high-precision triggers for class C are retained and the corresponding topic terms are added to lexicon L.

We also use a *junk* lexicon which contains stop words and low-information words (e.g. "lol", "omg", etc.).

Deals identification results: For this task, the feature set consists of tweet-level features (% tweet tokens from a *Deal* lexicon constructed as described above, % tweet tokens from the Junk lexicon, %tokens which are urls, etc.). Table 2 contains experimental results in two different settings. Using a *balanced* gold standard (200 positive elements, 200 negative elements) and a 10-fold cross validation setup, we obtain 95% precision and 93% recall. When experimenting with an *imbalanced* gold standard whose ratio of negative to positive examples reflects the results of our manual analysis in Table 1, the precision is 96% but the recall drops to 76%.

Events identification results: The setup for this experimental task follows that used for deal identification. On a small balanced dataset, events are identified with 96% precision and 97% recall, while on a larger dataset whose label distribution reflects the percentage of events in Table 1, events are identified with 80% precision and 84% recall.

Task	Gold standard	P	R	F-1
Deals	200+,200-	0.95	0.93	0.94
Deals	200+,6860-	0.96	0.76	0.85
Events	200+, 200-	0.96	0.97	0.96
Events	200+, 2860-	0.80	0.84	0.82

Table 2: Results: Deal and Event tweets can be sucessfully identified

Deals: top features	Events: top features
1. % tweet tokens from Deal lexicon	1. % tweet tokens from Event lexicon
2. % numeric tweet tokens	2. tweet length
3. % tweet tokens from Junk lexicon	3. % numeric tweet tokens
4. % tweet tokens which are URLs	4. % tweet tokens from Junk lexicon
5. tweet length	5. % tweet tokens which are pronouns

Table 3: Top ranked features for deal and event identification

Topic class	Num. topics	Example topics (manually labeled)
not clear	15	—
Calendar Occassion	11	New Year's, Valentine Day
Temporal Expressions	8	Hours, Week Days
Food and Drink	7	Chocolate, Ice cream, Beer, Wine
Sports	4	Superbowl, Spring training, March Madness
Weather	3	Snow, Rain
News	3	State of Union, Detroit/Ford news

Table 4: Examples: Business tweet topic classes

Month	Topic	Example terms
01-2010	Haiti Earthquake	haiti,relief,donate,cross,red,text
01-2010	New Year's	year, resolution, healthy, calories
02-2010	Valentine Day	day, valentine's, romantic, dinner,reservations
02-2010	Superbowl	super, bowl, sunday, saints, colts, party
03-2010	Saint Patrick's Day	day, st, patrick's, patricks, irish, paddy's, parade
03-2010	Drink specials	happy, hour, half, earth, drinks

Table 5: Results: Examples of business tweet topics.

In conclusion, events and deals can be identified with encouraging results. Table 3 contains the top 5 features ranked by importance for the two tasks.

4. BUSINESS TWEET TOPICS

In order to understand the *topics* of business tweets, we used the LDA package [1] to mine potential topics from 3 months of business tweets (examples are included in Table 5). We manually analyzed the top 25 topics from each month (75 topics total) and found that 60 topics (80%) were understandable and easy to describe. Table 4 contains manually derived *classes* of topics: in addition to those directly related to the business domain (e.g., Food, Drink), we find other general-interest topics such as holidays and celebrations, important political or sports events. Such topics reflect the way in which businesses adapt their communication based on the broader societal preoccupations and could potentially be used for better organizing and recommending business account content.

Reference

[1] A. Smola and S. Narayanamurthy. An Architecture for Parallel Topic Models, In *Proceedings of VLDB*, 2010.