$ale Cloud

Mid-Term Report

Team

Vimal Kini - UX, Coding, Research
Divya Karthikeyan - Project reports, Coding, Quality Analysis
Gaurav Nitin Shetti - Coding, Research, Product Management
Arian Shams - Project Manager, Coding

Mentor

Miguel Rios

Project goal

Visualizing sales/deals over Bay area and SF on the basis of Tweets. The project would encompass

1. Listening to the streaming API to detect tweets about sales/deals from SF/East Bay
2. Visualizing tweets related to sales on map as a “sales cloud” to allow users to instantly identify hot sales/deals that other shoppers are tweeting about.
3. Updating the map in real-time (about every ½ hour)
4. Allowing users to search by topics within the tweets and adjust the map accordingly. For instance, a user can search for tweets related to “Shoes,” which would then modify the map to show tweets relating to sales with the term “Shoes” in the text.
5. Allowing users to focus on certain regions of the map and view the sales tweets from those regions.

Project Timeline, Milestones and Strategy

Date	Task	Deliverable
Oct 26 **(Completed)**	Finalize Project Scope.	Setup project website
Nov 2 **(Completed)**	Complete research relating to filtering tweets against a certain topic (sales for us). Also, finalize front end user interface prototype.	Front end prototypes and summarization of research
Nov 9 **(Completed)**	Complete analysis of tweets relating to sales and develop algorithm for filtering sales related tweets.	Preliminary algorithm for filtering tweets
Nov 13 **(Completed)**	Have a functioning front end interface that is tested and works with the defined data structure that will feed into the interface. Also, create a preliminary design of how the backend servers will automatically capture tweets, filter them and then transform them to a structured data type that will feed into the front end.	Working site that maps data to an interactive map. Preliminary design of backend server setup
Nov 23	Finalize testing of algorithm that filters tweets relating to sales.	Finalized sales filtering algorithm.
Nov 30	Finalize font end user interface. Finalized implementation of backend servers that will capture, filter and transform sales related tweets to the proper data type to feed into the front end.	Functioning website that meets criteria listed in project goals
Dec 10	Quality testing and modifications (if needed). Complete write-up.	Project Complete

Grading Criteria

A = We have met all or all but one of the criteria listed in the Project Goals. The filtering algorithm that filters for sales related tweets has a reasonably low error rate and the front end interface is easy to use and provides value to the user. Most important of all, the final deliverable helps people find sales based on tweets in real-time.

B = We have not met two or three of the criteria listed in the Project Goals. The filtering algorithm has a high error rate (but is reasonable). The final deliverable helps people find tweets related to sales but many (if not most) of the tweets are not sales related.

C = We have not met four or more of the criteria listed in the Project Goals. The filtering algorithm has a high error rate that is unreasonable and does not help provide the proper tweets for the user to identify sales. The front end interface is not easy to use and does not provide value to the user. The final deliverable does not help the user identify sales due to unreasonably high error rates in filtering tweets and a poorly (or non-functioning) front end user interface.

Next Steps

For our next steps we plan to follow the project outline as shown previously. Our main concern is the volume of tweets that we will be able to extract that both are geo-tagged and have sale/deal related terms. Our advisor expressed this concern to us by stating that a very low number of tweets contain actual geo-tagged data and the search API does not provide accurate geo-tagged data since it may use a users' default location as stated in their profile. The main API feature for extracting geo-tagged data is from the streaming API. As a result of the low geo-tagged twitter data, we may adjust our "map" to be one that maps sales and deals by topic in some mapping visualization using tools such as d3.

Current Coding Files: https://github.com/gauravshetti/salecloud

Work Percentage

Tasks	Arian	Gaurav	Vimal	Divya
Literature Review	25	25	25	25
Twitter API Research	5	15	30	50
UI	70		30	
Backend		60	15	25
Report	50			50
Presentation		50	50	

Literature Review

The research for our project can be broken down into three main categories. The first category has to do with the search terms used by users where tweeting about e-Commerce related topics such as "sales" and "deals." The idea is to identify within some range of accuracy the various terms that users may use to describe tweets correlating with sales related events and then to filter or search for tweets that meet such criteria. The second category has to do with techniques for analyzing and mapping geo-tagged tweets. Specifically, we wanted to know how best to cluster geo-tagged tweets and to identify certain patterns within a geographic area in context to businesses located in that area. The third category of our research has to do with analyzing tweets, specifically in the context of lexical analysis, search and filter techniques, tweet categorization, and sentiment analysis.

eCommerce Search Terms

- A research conducted on users in the age group of 13 to 24 resulted in information about sharing trends among various social networking site users. [1]

- 10 most popular activities on Social networking sites: Looking at profile, Updating personal profile, searching for someone, emailing someone, writing on someone else's profile, reading blogs, listening to music, requesting friendship and looking up someone's status. Of these 2 activities, Updating personal profile and Looking up someone's profile are two important activities that aid in gathering information about how information about sales/deals is being diffused across users. [1]

- Among Twitter participants, 20% of status updates are specifically sharing information about brands leading us to consider specific stores or brands as a search term during the Tweet extraction. [1][13]

- Organizations use Twitter as part of their marketing strategies: Twitter as response mechanism - complaints or queries (if lower number of tweets, then reach and efficiency is lower).The efficiency can be increased by retweeting other's tweets and using mentions in tweets. Eg. Microsoft does this, which would steer us towards including Retweets in our analysis [6]

- Price, discount rate, category matter to the user who shares a daily deal. 70% of tweets shared via Twitter are in the range of 40-60% discount. Also deals present at multiple locations of a store are more likely to be shared. These aspects could be included during the Tweet filtering process to gather more specific data. [9]

Geo-Location Techniques

Measuring geographical regularity of local crowd behaviors by plotting the pattern of these 3 main indicators.

a. How many tweets are posted?
   #Tweets: the number of tweets that were written in an RoI within a specific period of time.
b. How many users are there?
   #Crowd: the number of Twitter users found in an RoI within a specific time period.
c. How active are the movements of the local crowd?
   #MovCrowd: the number of moving users related to an RoI within a specified period of time.

The measure of the number of Tweets at a geographical area along with determining the normal crowd behaviour would be beneficial in detecting an event. [2]

- Much less % of the microblog documents are geotagged. Hence determining a method to identify the location on non geotagged documents which is proven to increase the result by 115 times. [5]

- Challenges with geotagging in Twitter:

    a. Error rate of 967km
    b. Since Twitter documents are short, it may be impossible to predict the location solely based on terms

Alternative solution: Filtering method to detect documents with common theme allocated to the same place. [5]

- Geotag Allocator:

    a. Create a database of place names
    b. Place names are extracted from Twitter
    c. Compare both to find a similarity. Smaller the variance, more precise the name is
    d. If there is large variance (Eg: McDonalds is very ambiguous). Such entries can be eliminated from database

This could help determining if Tweets are sales based form a specific geographical area even though they may not be Geotagged. [5]

- Using Geolocation in Tweets to analyze large groups of people. Analyzing 2 types of tweets to gather geolocation data: (1) specifically geo-tagged tweets and (2) text only tweets where they had to extract the lat/long (geo-code) from the tweet by translating the place names using Google's geo-naming service. This could help determining if Tweets are sales based form a specific geographical area even though they may not be Geotagged. [7]

Tweet Analysis

- User/tweet categorizations (i.e. by gender, by geographic region, by tweeting habits, etc.) can be determined by taking a look at "trending" topics via the #hashtag in the tweets. UID as a measure of time as opposed to the datetime of the tweet. Could be useful to note as a way of filtering for tweets within a certain time period. The inclusion of Hashtags as a Tweet filter measure would be profitable while gathering sales related tweets [3]

- The effectiveness of Twitter data as a source of BI systems may critically rely on how well structural information on Twitter is exploited and how novel text mining techniques can be applied to analyze tweets.

    1. Weak ties are more likely to lead to retweeting
    2. For an advertising tweet, check if it contains an url
    3. Use a dictionary lexicon to further filter out the tweets
    4. 4 Categories of tweets helps: intention, positive, negative and neutral

Could consider if Url linking to specific deals could be included in the filter algorithm.[4]

- Metrics designed to capture text based opinion divergence in product reviews, adventure into an unexplored frontier in sentiment analysis. Their impact on consumer purchase behavior and product sales has significant implications for Word-Of-Mouth driven marketing research. [8]

- Among Twitter participants, 20% of status updates are specifically sharing information about brands leading us to consider specific stores or brands as a search term during Tweet extraction. [9]

- Traditional NLP tools do not work well with twitter data. Hence, utilize a named entity tagger trained on in-domain Twitter data presented in previous work. First annotate a corpus of tweets, which is then used to train sequence models to extract events. [10]

- There are three factors to look for when trying to extract events and event descriptions from Twitter. [12]

1. Main entity extraction: Use a large corpus and see manually the weightage of each word. If the inverse frequency is higher than other words, then it assumes higher importance and thus is the main entity. From than we can take the inverse document frequency.
2. Extracting actions and opinions: entity followed by a verb or noun phrase
3. Audience opinion extraction: Verbs preceded by a main entity or a pronoun followed by a verb phrase and then the main entity.

- Much of words in tweets are “ill-formed.” For example, one can write “tomorrow” as “2morrow.” There is a certain method that can be used to transform an “ill-formed” word (T) to a single “standard formed” word (S). There are 3 steps to this process: [11]

1. Identify which words are out-of-vocabulary (OOV). We can use GNU Aspell Dictionary to identify OOV’s. The categories of OOV’s include: Letter&Number, Letter, Number Substitution, Slang and Other [11].
2. Identify which words are ill-formed in relation to the string context. There are several methods where we can extract bi-grams to establish context. [11]
3. Candidate selection where we select an S for each T. There are a variety of algorithms that can be used for this.

This approach of lexical normalization may be cumbersome for our purposes since we are searching for only a few terms and are not in a need to normalize a large volume of tweets. The techniques used in this paper can be helpful for identifying OOV’s.

Bibliography

1. Park Y., Jaimie, and Chung, Chin-Wan. "When daily services meet Twitter: understanding Twitter as a daily deal marketing platform." *WebSci'12 3rd Annual ACM Web Science Conference.* Pages 233-242. 2012. http://dl.acm.org/citation.cfm?id=2380748&bnc=1

2. Burton, Suzan et al. "Interactive or Reactive? Marketing with Twitter." *Emerald Group Publishing Limited.* 2011. http://www.emeraldinsight.com/case_studies.htm/case_studies.htm?articleid=17003316&show=html

3. Cheong, Marc, and Lee, Vincent. "Integrating web-based intelligence retrieval and decision-making from the twitter trends knowledge base." *SWSM'09 2nd ACM Workshop on Social Web Search and Mining*. Pages 1-8. 2009. http://dl.acm.org/citation.cfm?id=1651439&bnc=1

4. Rui, Huaxia, and Whinston, Andrew. "Designing a social-broadcasting-based business intelligence system." *ACM Transactions on Management Information Systems.* Volume 2 Issue 4. 2009. http://dl.acm.org/citation.cfm?id=2070713&bnc=1

5. Watanabe, Kazufumi, et al. "Jasmine: a real time local-event detection system based on geolocation information propagated to microblogs." *CIKM'11 20th ACM International Conference on Information and Knowledge Management.* Pages 2541-2544. 2011. http://dl.acm.org/citation.cfm?id=2064014&bnc=1

6. Lee, Ryong, and Sumiya, Kazutoshi. "Measuring geographical regularities of crowd behavior for Twitter-based geo-social event detection." *LBSN'10 ACM SIGSPATIAL International Workshop on Location Based Social Networks.* Pages 1-10. 2010. http://dl.acm.org/citation.cfm?id=1867699.1867701&coll=DL&dl=ACM

7. Wakamiya, Shoko, et al. "Crowd-sourced urban life monitoring: urban area characterization based crowd behavior patterns from Twitter." *ICUIMC'12 6th International Conference on Ubiquitous Information Management and Communication.* Article No 26. 2012. http://dl.acm.org/citation.cfm?id=2184784&bnc=1

8. Zhang, Zhu, et al. “Deciphering word-of-mouth in social media: Text-based metrics of consumer reviews.” *ACM Transactions on Management Information Systems.* Volume 3 Issue 1. 2012. http://dl.acm.org/citation.cfm?id=2151163.2151168&coll=DL&dl=ACM

9. Jansen, Bernard, et al. “Being networked and being engaged: the impact of social networking on ecommerce information behavior.” *iConference’11.* Pages 130-136. 2011. http://dl.acm.org/citation.cfm?id=1940779&bnc=1

10. Ritter, Alan, et al. “Open domain event extraction from twitter.” *KDD’12 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* Pages 1104-1112. 2012. http://dl.acm.org/citation.cfm?id=2339704&bnc=1

11. Han, Bo, and Baldwin, Timothy. “Lexical normalization of short text messages: makn sens a #twitter.” *HLT’11 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies.* Pages 368-378. 2011. http://dl.acm.org/citation.cfm?id=2002520&bnc=1

12. Popescu, Ana-Maria, et al. “Extracting events and event descriptions from Twitter.” *WWW’12 20th International Conference Companion on World Wide Web.* Pages 105-106. 2011. http://dl.acm.org/citation.cfm?id=1963246&bnc=1