

Tweetviz: Visualizing Tweets for Business Intelligence

Bas Sijtsma
University of Amsterdam
Amsterdam, The Netherlands
btsijtsma@gmail.com

Pernilla Qvarfordt Francine Chen
FX Palo Alto Laboratory, Inc.
Palo Alto, CA
{pernilla,chen}@fxpal.com

ABSTRACT

Social media offers potential opportunities for businesses to extract business intelligence. This paper presents Tweetviz, an interactive tool to help businesses extract actionable information from a large set of noisy Twitter messages. Tweetviz visualizes the tweet sentiment of business locations, identifies other business venues that Twitter users visit, and estimates some simple demographics of the Twitter users frequenting a business. A user study to evaluate the system's ability indicates that Tweetviz can provide an overview of a business's issues and sentiment as well as information aiding users in creating customer profiles.

CCS Concepts

•Human-centered computing → Visualization design and evaluation methods; *User studies*;

Keywords

Social Media Analytics; Information Visualization; Location Profiling; Sentiment Analysis

1. INTRODUCTION

Today's widespread use of social media offers new opportunities for businesses to extract business intelligence. Consumers freely share their opinions about products and services at a large scale on platforms such as Facebook and Twitter. This provides a valuable resource that businesses can leverage for a competitive advantage. Marketers can mine the vast amount of data to detect and discover new knowledge, such as insights into customer interests, competitors offers, or detecting problems with a service, and use these insights to realize value and competitive intelligence. Traditionally, businesses spend a lot of effort obtaining customer opinions through focus groups, interviews, and so on. Hence, social data monitoring tools that assess the consumers' opinion are of major interest to businesses that recognize the benefits of social media.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

SIGIR '16, July 17-21, 2016, Pisa, Italy

© 2016 ACM. ISBN 978-1-4503-4069-4/16/07...\$15.00

DOI: <http://dx.doi.org/10.1145/2911451.2911470>

Twitter is a social networking platform where users can send 140 character messages ('tweets'). Each day its users' generates over 500 million messages¹. The large quantity of data necessitates automated tools to filter, order, mine and visualize it. In addition, the unstructured and noisy nature of social media data makes it difficult for companies to identify actionable areas of improvement.

We present Tweetviz, an interactive social media analytics system for exploring large collections of geo-tagged Twitter data. Tweetviz was developed based on requirements obtained from interviews with business owners and social media marketers. Tweetviz addresses two major problems: 1) identifying potential issues at a company's location; and 2) finding information about visits to competitors among their customers. Tweetviz is designed to provide location-specific actionable information that can be used to improve a business.

2. RELATED WORK

A number of commercial social media tracking products such as Hootsuite² and Sproutsocial³ focus on offering customer interaction statistics. They track metrics related to the number of responses to a company's messages, attempt to define the most successful response length, and other types of aggregated daily activity statistics. While this information is valuable from an on-line marketing perspective, this does not help businesses understand their customers' expressions at the venue level. Other systems such as Brandwatch⁴ and Zignal Labs⁵ focus in real-time event identification, e.g., identifying unhappy customers. Again, information at the venue level is not supported and analysts need to supply keywords of interest to provide to the system.

Several research systems have been developed for mining Twitter data. Diakopoulos' et al. developed a tool for mining current events from Twitter [2] which aimed to support journalists who wish to extract news from aggregated social media data. The user interface was specifically designed to allow for journalistic investigation of real-time responses to news events. Similarly, TwitInfo [3] allows users to explore real-time events occurring on Twitter. Both these systems are time-line based, and extract noteworthy elements based on tweet volume peaks and word-frequency based heuristics

¹<http://www.twitter.com/about>

²<https://hootsuite.com/>

³<http://sproutsocial.com/>

⁴<http://www.brandwatch.com>

⁵zignallabs.com

within a small time frame. Tweetmotif [4] is an application for extracting topics and summarizing sentiment from Tweets. This system allows exploratory search of any query given by the user. All three systems contain elements that are potentially valuable for a business analyst. However, none of these systems consider the geographical origin of messages, thereby losing a substantial level of context that could be used to gather business intelligence. The goal of this research is to leverage the geographical information to provide actionable location specific information.

Conversely, Chen et al. [1] mined social media data to profile businesses at specific locations. They matched geo-tagged tweets to Foursquare venues by performing density-based clustering to identify a tweet's specific business location. Then, the average sentiment of tweets at each location was computed, thereby providing a sentiment profile of the business. Tweetviz provides an interactive visualization for mining this processed tweet data together with information extracted from related social media data.

3. TWEETVIZ

The aim of Tweetviz is to provide business owners with actionable information about business locations. To understand what information business owners desired, we conducted interviews with business owners involved in decision-making or social media management. Based on the interviews, we identified three issues these stakeholders have when analysing Twitter feeds: 1) It's difficult to prioritize issues because of the large amount of data generated. Is the opinion expressed on social media shared by a large number of customers? 2) Identifying problems as well as positive aspects during a customer's store experience helps businesses identify concrete improvements to their service; 3) Knowing who their customers are and what they like can help companies tweak their service and target the right customer.

Based on our stakeholder interviews, we identify two design requirements. First, Tweetviz should support the analyst to discover the most important topics discussed at their business, both positive and negative. Second, the system should help the analyst obtain a profile of their customers and their behavior. This must be supported regardless of the volume of messages available.

3.1 Data used in Tweetviz

To support the functionality provided by Tweetviz, several datasets were collected. To present venue-level issues, the set of geo-tagged Twitter messages and automatically associated Foursquare business locations (i.e., *venues*) from [1] were used. The dataset contained around 24 million geo-tagged tweets from the San Francisco Bay Area, collected between June 4th, 2013 to March 23rd, 2015, with 656,098 distinct Twitter users. The dataset also included the estimated sentiment of each tweet, as described in [1].

The geo-tagged Twitter messages in the dataset were also used for estimating a customer's home location based on time and geo-location of the customer's tweets. In order to protect customers' privacy, the granularity of the estimation was reduced to neighborhoods using Zillow and Flickr neighborhood shapefiles rather than street address.

3.2 User Interface Design

Tweetviz was developed as a web application for modern web browsers. The front-end was built using the Angular

JavaScript framework, supported by a back-end written in Node.js. All data is stored in a MySQL server. The system is designed to allow browsing of the available data with direct updates without reloading the page whenever an interaction is performed. Figure 1 displays a screen capture of the system.

3.2.1 Venue Selection and Filtering

One of our design principles was to allow users to both get an overview and details about business locations in the user interface. After selecting a company from the company dropdown (Figure 1, item 1), the user can view aggregated tweet statistics (item 4) at the city, neighborhood or branch levels (item 3). The user can group multiple branches together, or zoom into a single business to view its available Twitter data. For easier access of particular cities, neighborhoods or branches, an analyst can search (item 2) locations by name.

3.2.2 Tweet Sentiment at Business Locations

To support quick access and comparison of the sentiment at different branches, each branch is marked on the map with a color corresponding to the location's average sentiment (item 8). Tweets of selected cities, neighborhoods and/or branches, are displayed at the bottom of the user interface (item 11).

To assist in the identification of positive and negative aspects about the business' service repeatedly voiced on Twitter, Tweetviz displays tweets at selected venues grouped by keywords (item 10). Keywords were automatically identified as frequent nouns (each tweet was parsed using the Stanford NLP parser⁶).

When a keyword and sentiment is selected, the list of tweets (item 11) refreshes to only show tweets matching the selection. Since the sentiment classification is not perfect, a user can easily reclassify, with a single click, a tweet as the opposite sentiment, or mark it as irrelevant. This functionality allows an analyst to organize and improve the quality of their company's data in order.

3.2.3 Customer Demographics

An analyst can select the demographic information tab (item 6) to view summary statistics of the neighborhoods their customers live in to support customer profiling. This view shows neighborhoods ordered by most unique customers to selected venues and average home prices for the neighborhoods.

3.2.4 Visits to Competitor and Other Venues

A Tweetviz user can explore the other venues a business's customers tweet about (item 7). Although not all businesses are competitors, finding out what other businesses customers frequent can support creation of customer profiles.

To determine what other businesses a venue's customers visit, all tweets for each distinct customer are retrieved. Tweets not sent from a business location are discarded. Next, the number of visits to other businesses are extracted and aggregated. The system keeps track of both the unique number of customers visiting a business and the total number of visits. The results are filtered based on the list of selected venues (item 4).

⁶<http://nlp.stanford.edu/software/lex-parser.shtml>

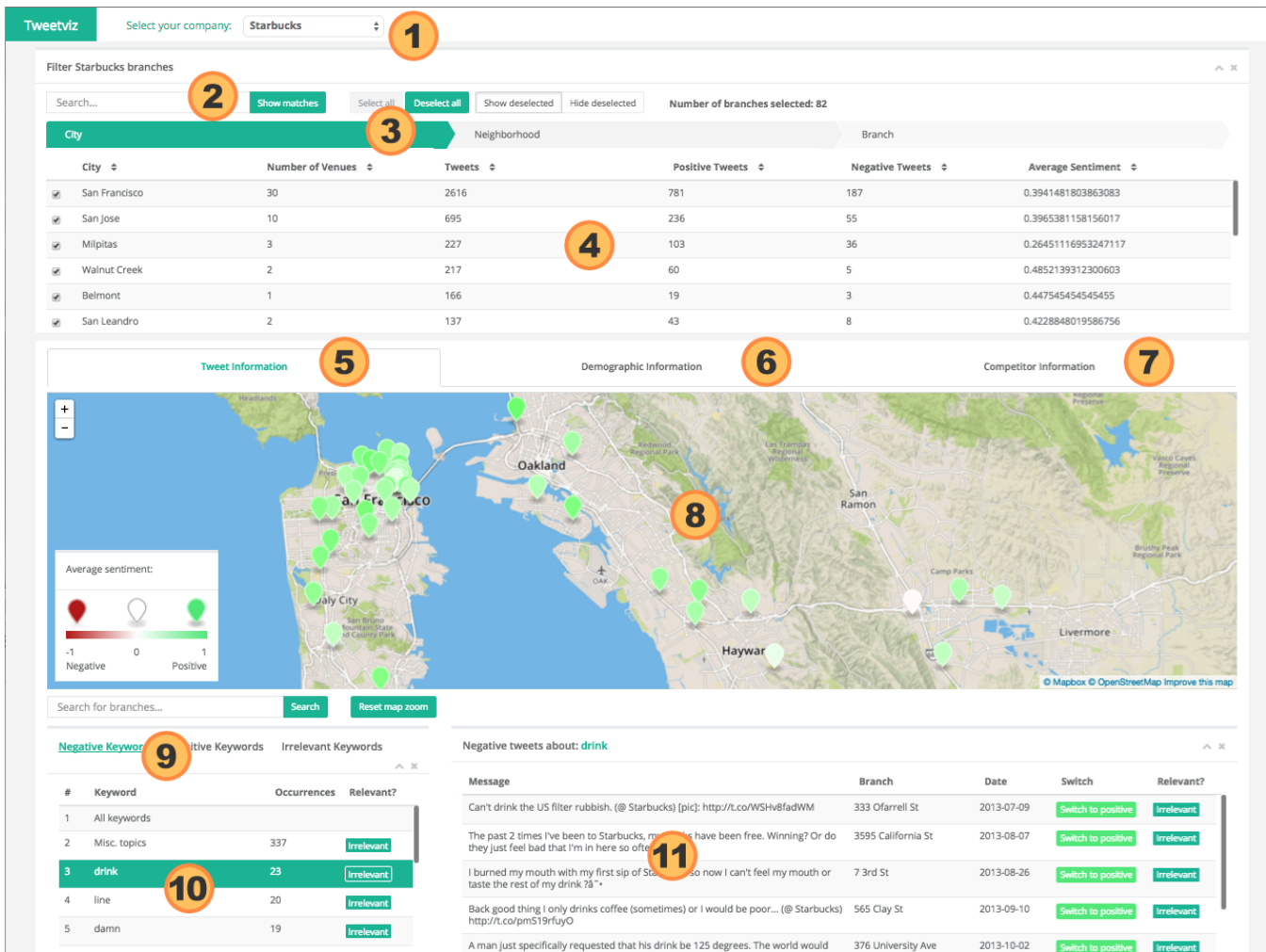


Figure 1: Screen capture of the Tweetviz user interface with elements marked with numbers in circles.

4. USER STUDY

We performed a user study to investigate users' experience of Tweetviz. We were in particular interested in understanding: 1) The system's ability to provide a valid overview of issues and sentiment voiced by customers; 2) The user experience; and 3) The type of inferences that participants draw given the available demographic and competitor information.

In order to assess the usefulness of automatic organization of noisy data by keyword, we compared Tweetviz as described above (aka experimental) with a version without automatic keyword identification (item 10 in Figure 1). In the baseline version, a list of all tweets for selected business were shown and the user could filter the list by typing a search query. Twelve participants used the experimental version of Tweetviz and another twelve participants used the baseline version. Of the 24 participants, 5 were women and 19 men and their ages ranged from 21 to 45 years. All the participants had experience of Twitter, and 20, or 83.3%, stated that they felt familiar or very familiar with Twitter.

Two types of tasks were designed: identification of business-related issues based on tweets (in total 4 tasks; 2 tasks using positive tweets and 2 using negative tweets) and profiling

a business' customers using the available demographic and competitor information (in total 1 task). The experiment was designed to represent real scenarios a business analyst may be working towards, as indicated by interview subjects.

Ground-truth for the issue identification tasks was created by retrieving all tweets relevant to a task, and manually grouping the tweets based on negative and positive issues mentioned in the tweets. Keywords were ranked according to frequency, with the most frequent given a rank of 1.

Before the participants were given the tasks, they went through a training phase. During training, the participants were introduced to Tweetviz's features one at the time and asked to complete a simple task using the feature. Instructions were shown in a separate browser window located to the right of a browser window with Tweetviz. Next, the participants were given the five tasks one at the time starting with the issue identification tasks. The order of the issue identification tasks was balanced over all participants. The participants were free to choose how they worked with the tasks. After all tasks, the participants were asked about their impression of Tweetviz.

Table 1: TP-rate of ground truth top 1-3 issues

	Top 1	Top 2	Top 3
Baseline	0.625	0.500	0.368
Experimental	0.708	0.614	0.500

5. USER STUDY RESULTS

5.1 Issue Identification

For the first four tasks, identification of issues, we compared the performance between Tweetviz with (experimental condition) and without (baseline condition) automatically generated keyword classifications.

5.1.1 Task Completion Time

The total time spent on all four tasks for the experimental condition ranged from 556.5s to 1247.4s, with a mean (μ) of 947.5s and standard deviation (σ) of 209.3; versus the baseline condition with ranged from 519.7s to 1390.3s, $\mu = 1017.1s$ and $\sigma = 253.6$. Although on average the tasks were completed faster on the experimental interface, an independent samples t-test revealed that the difference is not statistically significant ($t(22) = 0, 734, ns$).

5.1.2 Issue Identification Performance

The issues identified by the the participants were compared with the ground truth and assigned equivalent rank to matching ground truth issue. We found that the top three issues were more frequently found by participants using the experimental version of Tweetviz. Table 1 shows TP-rates for correctly identified issues with rank 1 (Top 1), rank 1 and 2 (Top 2) and top 3 ranks (Top 3). The TP-rate for Top3 was significant higher for participants using the experimental version compared with the baseline version ($t(90)=2.6455, p<0.01$). Many participants (7 out of 12) in the baseline condition requested that the tweets be better organized: *“It would be nice if there were a feature that grouped together tweets with similar semantics to make it easier to identify the top concerns or praise of customers, rather than requiring the user to examine every individual tweet when trying to find common trends in the data.” (P3)*.

5.1.3 Viewing Keywords and Searching Tweets

The two different versions of Tweetviz provided different functionality for parsing the tweets. In the baseline version the number of searches indicates that the search functionality was not well used, with $\mu = 1.0$ ($\sigma = 1.22$) searches performed per participant and task. A likely explanation is that instead of searching for possible terms, the participants scanned through all the tweets to identify business related issues.

In comparison, in the experimental version, tweets were automatically grouped by keyword. This functionality was heavily used by the participants; on average they clicked on 9.2 topics per task ($\sigma = 5.80$). We also found a strong significant correlation ($r = -0.7781, p < 0.05$) between the number of keyword clicks and issue identification performance. Hence, participants who clicked on more keywords found more of the top-ranked issues. This is confirmed by participants’ comments, such as: *“[Tweetviz] was good at interacting with a bunch of tweets at once” (P6)*. The implication is that showing tweets by keyword has the potential to be

valuable for discovering the most important issues from a large collection of tweets, even with noisy classification and noisy tweets.

5.2 Creating User Profiles

Although our participants had no experience in user profiling, when examining the results of the user profiling task, we found that some participants were able to create rich descriptions of the customers.

For instance, three participants tried to find the average distance customers were traveling to reach business locations using the map. They experimented with selection and deselecting branches in various parts of the region to determine which branches had the most customers who lived further away from. Examples of richer descriptions from the user profiling task were: *“[...] middle/high-class income, heavy coffee-user”(P8)* and *“Customers are trendy and like to spend time in coffee shops, the park, and shops at Whole Foods”(P11)* and *“They have some brand loyalty (more check-ins at Philz than any other coffee companies), but they do also check in at competitors, such as [...]”(P6)*.

These observations indicate that Tweetviz provides usable functionality to extract valuable information when effort is spent digging deeper into the data.

6. CONCLUSION

In this paper we present Tweetviz, a social media analytics tool designed to help businesses explore large volumes of geo-located Twitter data in order to extract actionable information. Tweetviz provides users with the ability to mine information about sentiment, possible competitors, and simple customer demographics at business locations. In a user study, we evaluated Tweetviz’s ability to provide a valid overview of business related issues and sentiment expressed by its customers. We demonstrated that by interacting with the automatically generated keywords, a users’ ability to discover the most important business-related issues in the noisy tweet data improves, and Tweetviz’s data visualization allow its user to generate rich user profiles.

7. REFERENCES

- [1] F. Chen, D. Joshi, Y. Miura, and T. Ohkuma. Social media-based profiling of business locations. In *Proceedings of the 3rd ACM Multimedia Workshop on Geotagging and Its Applications in Multimedia*, pages 1–6. ACM, 2014.
- [2] N. Diakopoulos, M. Naaman, and F. Kivran-Swaine. Diamonds in the rough: Social media visual analytics for journalistic inquiry. In *Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on*, pages 115–122. IEEE, 2010.
- [3] A. Marcus, M. S. Bernstein, O. Badar, D. R. Karger, S. Madden, and R. C. Miller. Twitinfo: aggregating and visualizing microblogs for event exploration. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 227–236. ACM, 2011.
- [4] B. O’Connor, M. Krieger, and D. Ahn. Tweetmotif: Exploratory search and topic summarization for twitter. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*, 2010.