The Dividends of Improving Best Practices for Social Media Research

JACQUELINE ANDERSON, GINA PINGITORE, PH.D., AND MIRIAM ECKERT

A J.D. Power and Associates White Paper

March 2012

Contents

Introduction

Social media is continuing to gain traction in the world of market research. While this is an exciting development that offers new opportunities for data collection, there is a potential downside. With so many tools available to so many different types of users, the "wild west" approach to social media data access may be leading us to questionable results. The ability to trust social media data that is delivered by market researchers will be called into question once stakeholders become aware of this issue. Without methodological rigor, results from social media research can at best simply be unreliable, and at worst lead decision-makers to draw wrong conclusions.

Social media is continuing to gain traction in the world of market research.

In this paper, we will show that without well-established and proven guidelines on query construction and data extraction, very different results and conclusions can be obtained by different analysts attempting the same social media data search. In extreme cases, analysts can create such highly divergent queries that the associated data leads to different answers to even simple questions, such as: Which brand is my main competitor? Is Product[1] more of my brand's conversation this month centered around product? Is the sentiment expressed toward my brand this month more or less positive than the sentiment expressed toward my brand last month?

The results of our findings suggest that if investments are not made to standardize these practices, the differences emerging in the data could have a significant impact on the credibility of market researchers using this method, as well as on bottom-line business decisions.

Extraction Methodology for Social Media Data

Social media data extraction tools are software that enables analysts to develop social media search queries for a particular brand, e.g., Brand Y, and a subtopic related to that brand, e.g., Price. The query is then used to extract relevant posts from an existing database of social media data. Such tools may also calculate consumer sentiment (positive or negative) expressed in each post toward the brand and each subtopic identified by the search query.

The search queries are typically relatively complex, as analysts need to capture all the different ways in which consumers may talk about a particular brand or topic. The query for Price, for example, may contain combinations of terms such as "cost," "rate," "expensive," "dollars," or "paying." Each analyst may develop a different strategy for creating queries, and there is a danger that the results may be strikingly different.

[1] Product is one of the J.D. Power 5Ps used in J.D. Power and Associates syndicated (industry-wide) studies: People, Presentation, Price, Process and Product.

In previous work (Pingitore, Eckert, & Li, 2011), we tasked six analysts with developing queries utilizing the NetBase Theme Manager tool for three different topics—Hotels, Airline Baggage Fees, and Telecom. Data extracted from each analyst's queries were compared against each other with respect to volume, net sentiment, and categorization accuracy.

Volume is defined as the number of posts extracted for a particular query for a given time period.

Sentiment refers to the positive or negative opinion expressed toward a given brand, product, or concept within a post. A post can be classified as positive, negative, mixed, or unknown. The sentiment for a particular query can be calculated in a number of different ways. The first is net sentiment, which is the ratio of negative to positive posts from a particular query in a given time period. Although frequently used by many companies, the notorious lack of reliability of most net scores (Pingitore, Morgan, Rego, Gigliotti, and Meyers, 2007) has prompted us to utilize other measures of sentiment. For example, sentiment can also be expressed by looking at the percentage of positive posts relative to the entire data set, or the percentage of negative posts relative to the entire data set.

Categorization accuracy refers to the percentage of posts that are truly relevant to the intended topic (e.g., "Sprint has great prices") and are not false positives (e.g., "Sprint Football Team member experiences the price of injury"). The categorization accuracy is calculated by having a trained annotation team label a sample set of the extracted data.

The results in Pingitore et al, 2011, which are shown in Table 1, exhibited a high degree of inter-analyst variation in the data for both volume and sentiment. The findings raise the strong possibility that any conclusions derived from the results could be radically different, depending on which analyst developed the query. For example, for the Telecom topic, Team 1 Rater 1's query extraction produced a higher number of posts for Price (753,318 total posts) than for Product (55,712), whereas Team 3 Rater 2's query produced more for Product (2,019,000) than for Price (571,206) for the same time period. Regarding sentiment, Team 2 Rater 1 resulted in a net sentiment rate of 3% for the Telecom Product topic, whereas Team 2 Rater 2 had more than five times that rate at 16%.

Table 2 shows the difference in measured accuracy among different analysts for the Hotel and Airline Baggage Fee topics. The accuracy for the Airline Baggage Fee topic ranged from 0% (Team 2 Rater 4) to 12% (Team 3 Rater 5). For example, this means that for one analyst's data, 12% of posts were not actually relevant to the intended topic and yet they would be counted as volume toward that topic if care is not taken to eliminate them.

Table 1: Inter-Analyst Variation in Volume and Sentiment Results across Different topics

		Hotels*		Airline Baggage Fees*		Telecom (People)*		Telecom (Product)*		Telecom (Price)*		Telecom (Presentation)*		Telecom (Process)*	
		Total Industry Volume	Total Net Sentiment	Total Industry Volume	Total Net Sentiment	Total Industry Volume	Total Net Sentiment	Total Industry Volume	Total Net Sentiment	Total Industry Volume	Total Net Sentiment	Total Industry Volume	Total Net Sentiment	Total Industry Volume	Total Net Sentiment
Team 1	Rater 1	1,627,577	9%	27,816	6%	56,448	3%	55,712	5%	753,318	6%	368,484	6%	213,294	19%
	Rater 2	1,941,923	9%	32,153	7%	127,799	4%	106,015	4%	1,457,505	6%	428,097	14%	15,009	4%
	ICC	0.97	0.95	0.92	0.9	0.79	0.58	0.79	0.75	0.78	0.94	0.94	0.7	0.08	0.45
Team 2	Rater 1	1,010,533	10%	23,718	9%	41,165	3%	104,857	3%	158,021	3%	104,379	10%	15,072	7%
	Rater 2	240,351	17%	4,397	3%	310,550	8%	248,760	16%	240,806	10%	66,550	9%	59,382	18%
	ICC	0.4	0.71	0.29	0.7	0.17	0.53	0.77	0.43	0.58	0.63	0.82	0.8	0.29	0.42
Team 3	Rater 1	2,176,050	8%	6,835	7%	22,124	5%	95,872	8%	98,778	7%	120,980	7%	64,863	16%
	Rater 2	1,873,525	9%	6,453	6%	553,722	5%	2,019,000	3%	571,206	11%	209,242	10%	1,039,261	10%
	ICC	0.91	0.9	0.95	0.9	0.04	0.96	0.47	0.4	0.22	0.86	0.7	0.78	0.09	0.43

* Averaged across brands; reliability varied at the brand level with some comparison larger and other smaller than the industry averages

** Team 1 had the most knowledge of Boolean logic and syntax

Source: J.D. Power and Associates 2011 syndicated studies

Table 2: Inter-Analyst Variation in Query Accuracy across Different Topics

		Hotels				Airline Baggage Fees			
		% Correct	% Info Sources	% False Positives	n	% Correct	% Info Sources	% False Positives	n
Team 1	Rater 1	48%	30%	12%	329	66%	25%	8%	329
	Rater 2	51%	36%	10%	342	60%	30%	7%	342
	Average	0.50	0.33	0.11		0.63	0.28	0.08	
Team 2	Rater 1	81%	32%	9%	632	89%	7%	3%	632
	Rater 2	36%	16%	21%	303	39%	59%	0%	303
	Average	0.59	0.24	0.15		0.64	0.33	0.02	
Team 3	Rater 1	56%	30%	12%	316	29%	55%	12%	316
	Rater 2	51%	36%	11%	423	33%	53%	11%	423
	Average	0.54	0.33	0.11		0.31	0.54	0.11	

Source: J.D. Power and Associates 2011 syndicated studies

The findings raise the strong possibility that any conclusions derived from these results could be notably different, depending on which analyst developed the query. For example, given inconsistent and fluctuating sentiment levels in the data sets, one analyst may come to the conclusion that Telecom Brand A has more negative consumer sentiment associated with it than does Telecom Brand B. A second analyst, however, may assume the exact opposite. Similarly, differences are also found in comparing volume of conversation such that one analyst's results suggest that there is more conversation about Price than there is about another term, "customer service." In contrast, another analyst's results might indicate that customer service is a bigger conversation driver than Price.

Improving Accuracy, Validity, and Reliability

Using a detailed error analysis report created by our quality control (QC) team, we have determined the primary factors responsible for the observed variation in the previous study. These include the use of sentiment expressions such as "love" or "annoy" in queries, and the attempted exclusion of news sources by some analysts. Expanding on our previous work, this paper establishes social media standards around query development that includes developing Boolean logic templates aimed at improving accuracy, validity, and reliability. Increasing these three hallmarks of measurement will help social media data become a more viable source of market research insights. Additionally, we determined the need for creating and utilizing a QC team that works closely with both the analyst and the client-facing team to ensure the extracted data addresses the business questions being researched. We show how these processes allow more confidence in the business solutions we can make based on the social media data.

In this paper we will focus on the Products topic in Telecom as a case study. We will compare the variety of different conclusions that could be drawn from the first research phase with the more consistent and accurate conclusions that could be drawn from the results of the second, refined research phase. We will then explore how the social media query development guidelines and best practices directly affect clients by preventing costly marketing and strategy errors, and how they also benefit the overall industry by bringing more reliability to the research around social media.

Data Analysis

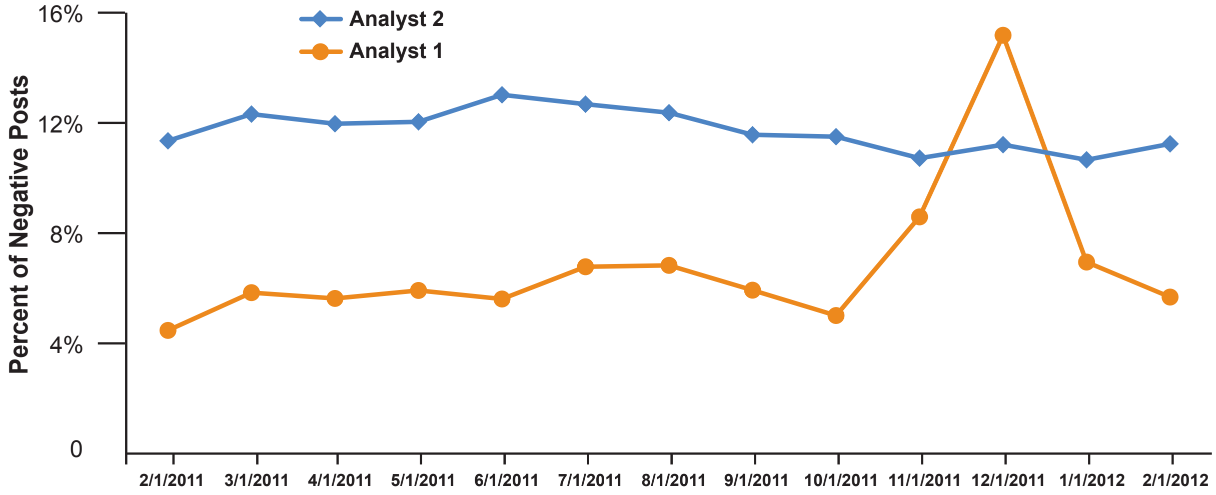
AN EXAMPLE OF VARIATION AT THE BRAND LEVEL

In our initial experiment, analysts were tasked with creating a query to extract social media posts written by consumers about a particular wireless brand (Pingitore et al, 2011). Based on the results, we developed a set of best practices for query development that included the following:

1. Be specific in defining your topic
2. Establish the right balance between precision and recall
3. Avoid sentiment expressions in queries
4. Employ well-trained analysts
5. Utilize separate QA teams
6. Ensure proper feedback[2]

After developing a set of guidelines for query development and implementing a QC process, an experienced analyst created a separate query for the same topic. The table below compares the results of an analyst in the first group (Analyst 1) with those of an analyst in the second group (Analyst 2) who had been instructed to adhere to the best practices and whose results had been reviewed by the QC team. The results are strikingly different. Analyst 1's data showed a sharp increase in negative sentiment posts in December 2012, whereas the curve of the data set of Analyst 2 remains relatively flat.

Brand X Brand Level–Percent Negative Sentiment Posts (2/2011-2/2012)



Source: J.D. Power and Associates

Figure 1

Examination of Analyst 1's data showed that the increase is largely due to mentions of a merger of Brand X and Brand Z falling through, which was announced in that month and which became a significant news event.

[2] See Pingitore et al, 2011, for a precise description of these recommendations.

QUERY DIFFERENCES AT THE BRAND LEVEL

A comparison of the queries and the extracted data revealed two main differences:

1. Analyst 1 did not successfully exclude non-consumer-generated posts, i.e., posts that were not written by actual wireless users, whereas Analyst 2, using the new guidelines, included a clause that eliminated a large portion of news and informational data.
2. Analyst 1 included the term “mobile” as an alternative expression for cell phone in the query. The QC process revealed that this led to the inclusion of mentions of a different brand, Brand W. Analyst 2, on the other hand, included more specific terms, such as “mobile service,” “mobile provider,” and “mobile carrier,” which allowed for the same coverage but successfully excluded mentions of the irrelevant brand. When Analyst 2’s query was modified to include just the term “mobile,” a similar negative spike was observed.

These two query differences led to data differences that manifest themselves both in terms of **quantity** and **quality**. Quantity of negative posts is affected because large numbers of repetitive news posts and reposts are included in the volume. Quality is affected because Analyst 1’s query led to non-consumer-generated posts such as the following, which dominated December 2011 data and resulted in a negative spike:

1. *[Brand X] strategy annoys judge in [Brand W] case*
2. *[Brand X’s] talks of potential [Brand W] asset sale reportedly falter*

Analyst 2’s query, on the other hand, excluded such posts, and the negative December 2011 data, like that of the other months, consisted of consumer-generated posts, e.g.:

1. *[Brand V’s] wireless service sucks. I hate how I can’t get the unlimited data plan, but my mother can.*
2. *I really hate [Brand V]! My phone is only 7 months old and of course is broken, have to wait 4 to 6 days for a new one.*
3. *Kinda mad tho , in order for me to get another phone i gotta wait until my 90days are up.. [Brand V] sucks !*

Analyst 1’s query resulted in a further problem of negative mentions around Brand W being counted against Brand X, such as the following:

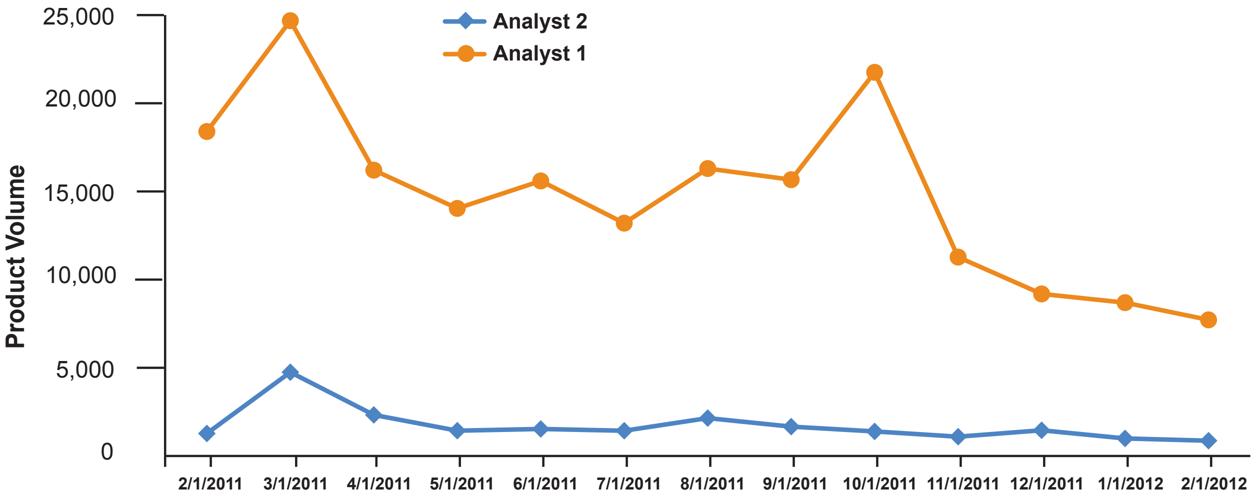
1. [Brand X] RT @Samxxx: I hate [Brand W] .. I wanna switch to [Brand U] ^_^.

An Example of Variation at the Subcategory Level

As part of the initial experiment, the analysts were asked to develop queries for the 5Ps within the brands (People, Presentation, Price, Product, and Product). Whereas brand-level queries are often relatively simple, subcategories intended to cover vague topics, such as customer service or quality of the product, require a great deal of research and creativity on the part of the analyst. When consumers talk about customer service, they may occasionally use that term (customer service), but more likely will talk about the rudeness of a particular employee, the long time they spent waiting on the phone, or the prompt attention they received upon entering a store. Similarly, when talking about the quality of a wireless product, consumers may address a whole host of issues, such as dropped calls, the fuzzy sound quality, the static, or the small screen on the handset.

The graph below shows the posting volumes for the Product (quality) subcategory of the Brand V brand from one of the initial queries (Analyst 1) and the posting volumes for the same query after implementation of guidelines and a QC process (Analyst 2). Interestingly, the volumes resulting from Analyst 2's query are much higher (ranging between nearly 7,000 posts and 25,000 posts) than those of Analyst 1 (ranging from a few hundred to 5,000).

Brand V Product Volumes (2/2011–2/2012)



Source: J.D. Power and Associates

Figure 2

Another difference is that the line for Analyst 2 shows much more variation, with two distinctive volume spikes occurring in March and October. The line for Analyst 1 does not mirror the October spike at all.

QUERY DIFFERENCES AT THE SUBCATEGORY LEVEL

Looking at the results of the Product queries of the two analysts, shown in Figure 2, the most striking distinction is that Analyst 1's query consists mainly of a small set of simple nouns and adjectives that cover the most basic and obvious quality mentions, such as “dropped calls,” “coverage,” “network quality,” and “static.” Analyst 2’s query, on the other hand, is far more complex and attempts to capture a much wider range of ways in which consumers may refer to the quality of the products. It includes complex phrases and Boolean expressions capturing the many different ways in which an actual consumer could refer to issues with the quality of the product, e.g., “acting up,” “crash,” “is down,” “hard to hear,” “cant’ get,” AND [email OR message OR voicemail OR text], “call,” AND “fail.”

This difference helps explain the significantly higher volumes of Analyst 2’s query, which has much greater coverage. Many of the relevant posts extracted by Analyst 2’s query would be missed by Analyst 1’s query, e.g.:

1. *Update, I can recieve text messages from [Brand U] and [Brand V] phones now. Still cannot recieve them from [Brand Y] phones.*
2. *Hey @[Brand V], I get an error message when I’m trying to upgrade to iPhone 4S. For some reason, [Brand Y] works fine.*
3. *I would like to propose a deal to [Brand V]. I will pay double my monthly cell bill if you will pay me back 5$ everytime I have a call fail.*

An examination of the sound bites from the October 2011 time period shows that in Analyst 2’s results, there are many mentions of quality complaints and praise around a new device launched by Brand Z, which was released during that month.

1. *I like my non-dropped calls and industry leading LTE speeds. Screw [Brand V] Meh. I’ll turn the 4G off as soon as I get my phone.*
2. *[Brand Z] 4 16GB ([Brand V]) 5.0 - Black - Works Great!!! http://t.co/6B6hQXRD.*

The limited query written by Analyst 1 has such low coverage that the jump in user posts is missed almost entirely.

Summary of Data Differences

We examined analyst differences in both brand- and Product-level findings of volume and sentiment and observed that relatively minor query differences lead to very significant changes in these important outcomes. For example, using the term “mobile” instead of “mobile service” and failing to account for non-consumer-generated conversation resulted in notably different results. Only a systematic checking of the sound bites from the initial query revealed that a spike in negative sentiment was due to non-relevant news posts about a different brand being included, as opposed to reflecting an actual shift in consumer sentiment.

At the subcategory level, we found that the two analysts’ queries also differed in terms of coverage. The vaguer nature of the subcategories requires much longer and more complex queries. Only Analyst 2’s query was inclusive enough to capture the many creative ways in which consumers may express themselves. As a result, Analyst 1's query missed the increased buzz resulting from a product release. However, such expansive and complex queries need to be submitted to a testing and revision cycle to ensure that categorization accuracy is at acceptable levels.

Business Implications

Understanding how subtle differences in query development can impact social media findings should make the issue of data quality an imperative for anyone using this data to inform business decisions. Revisiting our first example of the Brand X brand overall, had the market research (or PR or marketing) group used Analyst 1's data for an ongoing tracking study, they would likely have alerted key stakeholders to a potential brand crisis in the December 2011 time frame. The notable peak in negative sentiment might have caught the attention of everyone in the organization and may have prompted them to investigate the reason why the brand received such negative exposure. At the very least, this would result in many labor hours needlessly spent on investigating the data to find a solution (and, for an untrained analyst, would have been a very time-consuming process). At worst, the company could have produced a marketing blitz, trying to roll out new campaigns to counteract the negative sentiment they perceived existed. On the other hand, had Analyst 2's data been used, daily work could have continued as usual since there really was no concern to respond to. In fact, the data shows an almost opposite effect, with overall negative sentiment actually decreasing slightly.

In the second example, had Analyst 1's data been used, key consumer conversations would have been missed completely. Stakeholders would have missed the chance to insert themselves into the abundance of conversations (both positive and negative) that were occurring around the launch of a new product. These conversations were ripe with insight regarding opportunities to provide improved customer service, as well as opportunities to understand consumer reactions to the product itself. These conversations could even create opportunities for cross-sell of additional products such as accessories. By failing to uncover these conversations, the product development, marketing, and research teams would have missed valuable opportunities to not only learn more about their customers and products, but also to create valuable touch points with their customers.

Conclusion

Social media is increasingly becoming a valuable part of the consumer landscape, making social media an important source of insight for all market researchers. However, before we depend on social media as a guidepost for making critical business decisions, we have a duty to create more rigor around the processes that create social insights.

By taking the time now to research and develop best practices, we can create processes that will establish our ability to stand behind the data as it becomes an increasingly important part of the research mix. The investment of time and effort at this stage will ensure that market researchers are able to maintain their reputation of presenting stakeholders with only quality data and insights. Creating these standards will also ensure that research has a voice in the overall social media conversation within a company. Failure to create and adopt these social media data best practices leaves the future of social market research to the wolves.

Bibliography

Pingitore, G., Eckert, M., & Li, S. (2011, October). Understanding and improving the quality of social media data. *Research World*.

Pingitore, G., Morgan, N., Rego, N., Gigliotti, A. & Meyers, J. (2007) The single question trap. The Net Promoter score has limits in predicting financial performance. Market Research

The information contained herein has been obtained by J.D. Power and Associates from sources believed to be reliable. However, because of the possibility of human or mechanical error by these sources, J.D. Power and Associates does not guarantee the accuracy, adequacy, or completeness of any information and is not responsible for any errors or omissions or for the results obtained from use of such information. J.D. Power travel industry customer satisfaction studies (hotel, rental car, airline, and airport) were included in this special report. These studies include some responses from Canadian consumers.