THE IMPACT OF NATURAL LANGUAGE PROCESSING-BASED TEXTUAL ANALYSIS OF SOCIAL MEDIA INTERACTIONS ON DECISION MAKING

Larson, Keri, University of Alabama at Birmingham, BEC 317B, 1720 2nd Ave S., Birmingham, AL, USA, kerilarson@uab.edu

Watson, Richard T., University of Georgia, 316 Brooks Hall, Athens, GA, USA, rwatson@uga.edu

Abstract

Organizations typically use sentiment analysis-based systems, or even resort to simple manual analysis, to try to derive useful meaning from the public digital "chatter" of their customers. Motivated by the need for a more accurate way to qualitatively mine valuable product- and brand-oriented consumer-generated text, this paper experimentally tests the ability of an NLP-based analytics approach to extracting knowledge from highly unstructured text. Results indicate that for detecting problems from social media data, natural language processing outperforms sentiment analysis. Surprisingly, the experiment indicates that sentiment analysis is not only no better than manual analysis of social media data toward the goal of supporting organizational decision-making, but may even prove disadvantageous to such efforts.

Keywords: *text mining; natural language processing; sentiment analysis; social media*

1 The Promise of Natural Language Processing

Social media exchange is a now-ubiquitous mechanism through which consumers disseminate and elicit information in such forms as opinions, suggestions, and requests (Demetriou & Kawalek 2010). These consumer-generated data herald an increasingly valuable opportunity for organizations to create business value (Culnan et al. 2010; Hoffman & Fodor 2010). However, unearthing potentially valuable intelligence from these data also poses a significant challenge. Particularly, social information systems necessitate new tools for the real-time mining of huge volumes of *unstructured* text. For firms to detect important cues like adverse event mentions and consumer reactions to new products, analysts require the ability to qualitatively mine textual data. However, a notable gap exists between actual and desired capabilities for extracting latent information using existing tools. Therefore, a central question for social media researchers is whether a theoretically informed, natural language processing (NLP) approach to text-data analytics can confer an informational advantage to organizations over prevalent approaches currently available, particularly those based on sentiment analysis (SA).

A superior social media analysis mechanism is likely to exceed the simple positive-negative labelling capability of SA, in which a segment of text is categorized as positive, negative, or neutral based on word-level calculations (Pang and Lee 2004, 2008). Similarly, the simpler technique of counting characteristics like number of followers, number of likes, etc. is another important, but also incomplete, method for extracting knowledge from consumer-generated data. Despite the prevalence of SA as the basis of many social media brand reputation-monitoring tools (FreshNetworks 2011), we point out the large degree of meaning and knowledge lost by simply sorting suggestions, comments, and complaints into negative and positive piles. Recent studies demonstrate the range of distinction lost through such simple scales (e.g., Pavlou and Dimoka 2006). To illustrate, an extremely harmful problem may be indistinguishable from a mildly problematic observation if the treatment is limited to categorizing the notification as either positive or negative. Once extreme emotion is discerned according to a given lexicon, extracting the *subject* of the emotion requires further processing because SA is not able to understand the substance of a concern. Either a human reader must then manually interpret the comment to determine its significance (teams of whom organizations employ at great cost), or some type of machine-based algorithm must be further applied for qualitative analysis. This leads us back to the original requirement of a tool capable of *contextual* text data mining.

A viable candidate for extracting meaning from social media discourse, NLP blends computer science, machine learning, and linguistics in an aim to "understand" text in its natural format (Rajman & Besançon 1998, p. 51). NLP encompasses a wide range of disciplines and tasks focused on extending the capabilities of text mining, or the extraction of knowledge from unstructured text (Hearst 1999), most recently by incorporating the machine-learning (ML) paradigm of language processing. NLP algorithms have met with some success in formal, structured domains with limited lexes such as medicine and biochemistry (Tanabe et al. 1999). While the recent reinvigoration of NLP-related research is shepherding progress of machines to discover new, non-trivial knowledge from free text, the automated mining of data from unstructured text is still in its relative infancy. Emerging subfields and approaches continue to extend text mining proficiencies in the contexts of real-world data. Incremental improvements to a wide range of specific capabilities combine to contribute to discipline-level progress (Read 2005), and suggest potential applicability in less-structured or unstructured text environments such as social media (Bunescu and Mooney 2007; Kao and Poteet 2007; Agichtein et al. 2008). Advances include the automation of lexicon augmentation in named entity recognition (the accurate labeling of persons, organizations, and locations (Sang & De Meulder 2003)), parts-of-speech tagging, parsing (determining the grammatical tree of a sentence), and anaphora resolution (determining which noun or name a pronoun refers to).

We recognize a natural alignment between the knowledge discovery goal of NLP-based automated text mining and the organizational goal of extracting knowledge from highly unstructured customer exchanges. As a result, we are interested in determining whether NLP-based approaches might provide a decision-making advantage to firms, or whether existing manual or basic sentiment-based techniques are sufficient. In the case that our suspicions are incorrect, and current techniques adequately detect critical problems and opportunities from highly unstructured Tweets, updates, and comments, we would be in a position to inform both research and practice regarding the development of NLP-based social media analytics tools. We would thus conclude that efforts expended by computational linguists, artificial intelligence programmers, and computer scientists to develop machine understanding of unstructured text would be more usefully channeled into domains that, unlike social media, are characterized by *constrained* forms of text.

Intuition, however, leads us to suspect that NLP-based text data mining systems will in fact prove critical to firm-level decision-making in this day of pervasive, application-mediated textual discourse. Assuming the detection of significant problems and opportunities voiced by consumers can improve downstream decisions made by managers, it then follows that capabilities conveyed by an NLP-based social media analytics tool would benefit firms and consumers alike. Thus, continued investment of time and intellect would be warranted. In light of these consequences, this experimental investigation pursues the following research question:

Can advanced natural-language-processing-based qualitative textual analysis techniques improve the decision-making capability of organizations?

This paper empirically demonstrates that word-based sentiment analysis of social media text is no better than random sampling as the basis for grasping the problems that underlie customer chatter, and in fact may prove detrimental to a firm's attempts to accurately understand its customers' interactions. We demonstrate that an NLP-based approach can substantially enhance a firm's ability to detect problems with potential importance to its organizational decision-making from consumer chatter.

2 Text Mining and Sentiment Analysis

Text mining encompasses an array of theoretical approaches and methods likely to contribute to the ultimate success of machine-supported analysis of text (Feinerer et al. 2008). Such technologies include information retrieval, clustering, classification, entity-relationship and event extraction, and NLP (Hotho et al. 2005). The last is of particular interest to us in our quest to develop an approach robust against the idiosyncrasies of highly unstructured user-generated text. Relevant to our goals, the general objective of NLP is to create algorithms capable of "understanding" natural language through techniques ranging from the simple manipulation of strings to the automatic processing of natural language inquiries (Hotho et al. 2005).

In contrast to NLP's potential to unearth unknown, meaningful intelligence from text, organizations have relied on what is currently available—sentiment analysis. The immediate need to analyze volumes of social media data has forced firms to rely on a method that we argue is ineffective *in this setting*. While SA has the potential to assist insight into customers' reactions to products and services, we are dubious of its practical accuracy on an automated and large scale. We argue that any analysis of text at the word level, which necessarily ignores the aggregate meaning of whole clauses, is susceptible to misclassification. Idioms, negations, irony, sarcasm, slang, and misspellings—all prevalent characteristics of informal social media exchange—serve to obfuscate meaning. Even without such confounding elements, dissecting sentences into buckets of unrelated words decontextualizes each instance of each word. This step disengages a customer's intended meaning from the *assembly* of words he used to express that meaning. Mental-models research indicates that humans understand patterns of words locally, meaning that multiple instances of a single word situated among different surrounding words are not perceived by most English speakers as semantically related (Fox 1986). For example, we do not consider "my soup is *cold*" to have any relation to "I have a head *cold*." If we extract "cold" from the rest of the sentence in which it exists, which is equivalent to what happens during SA, we then have no idea what the word actually means. Thus, we are unable to determine whether it should be interpreted as a positive, negative, or neutral sentiment. Considering this critical loss of meaning inherent in any word-based approach, it is clear that contextual sensitivity is critical to a useful social media analytics system.

3 Propositions and Model

The major question we expect to answer through this investigation is whether an advanced, NLP-based analysis technique can improve the decision-making capacity of managers, specifically in the context of highly unstructured text generated by consumers within social media environments. We derive from our literature review a set of propositions whose outcomes will increase our knowledge of this domain.

Discussions with practitioners indicate that some organizations employ teams of social media analysts to manually sort and interpret customer comments. This manual method, while likely effective due to the application of human perception, interpretation, judgment, and reasoning, is conversely neither efficient nor cost-effective given the human element. Further, it is reasonable to expect information overload to occur once an analyst reaches the tipping point at which input (i.e., multiple streams of real-time social media data) exceeds processing capacity (reading and interpreting) (Miller 1956, Speier, Valacich, and Vessey 1999). Information overload reduces the quality of decision-making and increases both confusion and time to decision during the process (Speier 1999; Eppler and Mengis 2004). In short, information overload impairs decision-making. We are interested in facilitating the opposite.

In contrast to manual analysis of social media content, the automated approach underlying most systems is sentiment analysis. Despite its inadequacies as an analytical tool for highly unstructured text, roughly sorting negative from positive content may nonetheless serve a heuristic purpose, conveying a rudimentary refining mechanism that entails *some* degree of advantage over fully manual analysis of large datasets. Starting with a reduced set of pre-identified messages should ostensibly enhance problem/opportunity assessment by lessening the potential for cognitive overload. As such, we propose the following:

Number of detected critical problems and opportunities (SA-based)

***P1a**: Individuals assisted by sentiment-based machine analysis of social media content will detect a greater number of key problems than individuals with no machine assistance.*

***P1b:** Individuals assisted by sentiment-based machine analysis of social media content will detect a greater number of key opportunities than individuals with no machine assistance.*

Accuracy of detected critical problems and opportunities (SA-based)

***P2a**: Individuals assisted by sentiment-based machine analysis of social media content will more accurately detect key problems than individuals with no machine assistance.*

***P2b: Individuals** assisted by sentiment-based machine analysis of social media content will more accurately detect key opportunities than individuals with no machine assistance.*

Confidence in detection of critical problems and opportunities (SA-based)

***P3a**: Individuals assisted by sentiment-based machine analysis of social media content will have greater confidence that they detected key problems than individuals with no machine assistance.*

***P3b: Individuals** assisted by sentiment-based machine analysis of social media content will have greater confidence that they detected key opportunities than individuals with no machine assistance.*

Despite expecting SA to provide a small advantage, we doubt its capacity on a large scale to match the decision-making support capabilities of an advanced NLP-based system. We reiterate the potential loss of meaning with a word-based approach, in contrast to the contextualized evaluation of entire comments or clauses. The latter is far more robust to irony, sarcasm, misspellings, omitted words, and idiomatic expression, all inescapable characteristics of the highly unstructured text comprising social media communication.

Based on the information-detection advantage of an NLP-based approach, we propose the following relationships between NLP-based and SA or manual analysis of these data. Operationalizations of all propositions can be found in Table 1 in the Research Methodology section.

Number of detected critical problems and opportunities (NLP-based)

__P4a__: Individuals assisted by natural language processing-based machine analysis of social media content will detect a greater number of key problems than individuals assisted by sentiment-based machine analysis or with no machine assistance.

***P4b: Individuals** assisted by natural language processing-based machine analysis of social media content will detect a greater number of key opportunities than individuals assisted by sentiment-based machine analysis or with no machine assistance.*

Accuracy of detected critical problems and opportunities (NLP-based)

__P5a__: Individuals assisted by natural language processing-based machine analysis of social media content will more accurately detect key problems than individuals assisted by sentiment-based machine analysis or with no machine assistance.

***P5b: Individuals** assisted by natural language processing-based machine analysis of social media content will more accurately detect key opportunities than individuals assisted by sentiment-based machine analysis or with no machine assistance.*

Confidence in detection of critical problems and opportunities (NLP-based)

__P6a__: Individuals assisted by natural language processing-based machine analysis of social media content will have greater confidence that they detected key problems than individuals assisted by sentiment-based machine analysis or with no machine assistance.

***P6b: Individuals** assisted by natural language processing-based machine analysis of social media content will have greater confidence that they detected key opportunities than individuals assisted by sentiment-based machine analysis or with no machine assistance.*

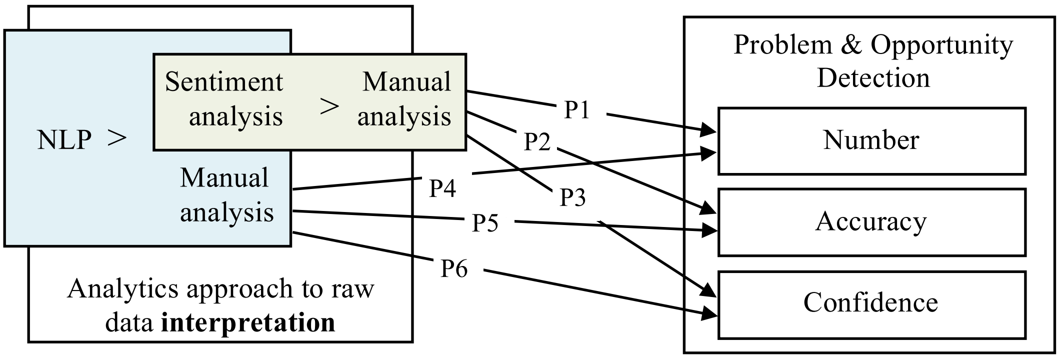


Figure 1. Research model highlighting propositions

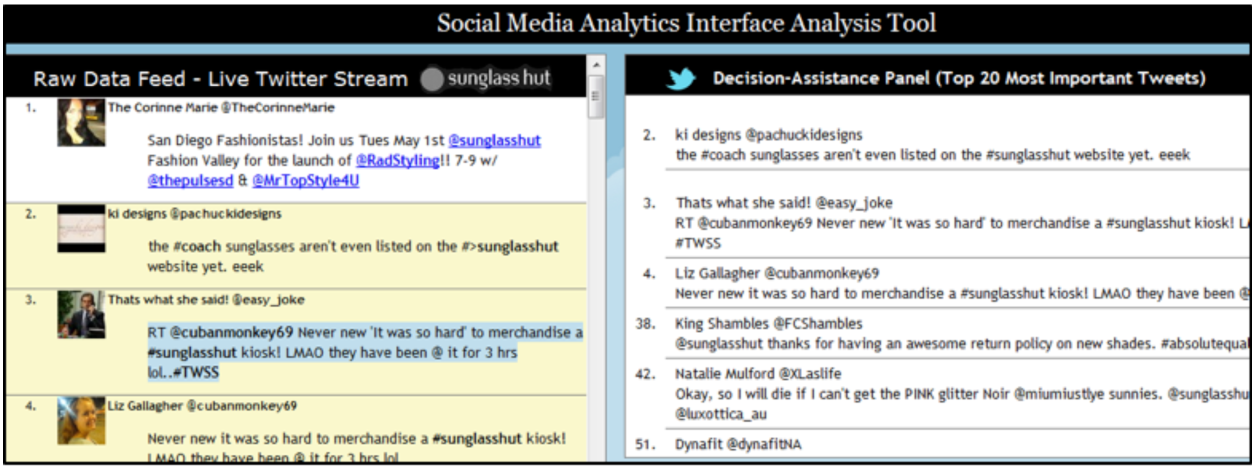
4 Research Methodology

The propositions were tested through a single-factor, controlled laboratory experiment. The experimental design included three *between-subjects* conditions representing *type of analytic approach* used to generate the decision-assistance panel: (1) *no analytical preprocessing*, (2) *SA-based preprocessing*, and (3) *advanced NLP-based preprocessing*. These conditions correspond with social media analytics approaches either currently employed in practice, or in developmental stages: (1) manual text monitoring, (2) the standard automated approach used in brand monitoring, and (3) a potentially useful innovation for firm-level social media monitoring, all of which are discussed in previous sections.

All subjects received three components via a browser-based interface designed for this experiment: (1) raw data (300 real Tweets in real order) streaming down the left half of the browser window, (2) a decision-assistance panel of 20 important Tweets pulled from the raw data according to treatment in the upper right quadrant, and (3) the debriefing questionnaire in the lower right quadrant. The experiment was implemented on a standard monitor in the same web browser to control for possible differential effects of look and feel. Font size, color, scrollability, etc., were hard-coded to ensure total uniformity across displays.

Participants in all conditions received raw data *identical in content and order* to ensure consistency of encountered problems and opportunities. The single difference across conditions was the decision-assistance panel; each treatment received a different set of 20 Tweets from the raw data, chosen according to the analytics approach being tested. We chose Twitter-mediated social media messages pertaining to the SunglassHut brand for the experiment because Tweets are restricted to 140 characters. This enabled us to control for maximum message length and by extension, density of information conveyed in a single message. SunglassHut is an ideal brand for the experiment because it is an operational business with a strong social media presence (resulting in abundant real raw data) that sells products accessible and familiar to our college student subject base. We confirmed through pilot testing that a stream of 300 Tweets provides sufficient information overload to prevent subjects from easily processing all messages manually, compelling subjects to rely on the decision-assistance panel provided to support task execution. This is a critical design feature since the ultimate goal of the experiment is to test the efficacy of *automated* preprocessing. Search functionality was also provided to simulate keyword searchability.

Part of the experimental interface is depicted in the following figure. A portion of the decision-assistance panel for the NLP treatment is show on the right. The 20 assistance Tweets are highlighted as they appear within the raw data stream (left) in yellow. This enables subjects to easily detect where in the raw data stream the assistance Tweets occur, thus ensuring context, which is important when Tweets build upon or respond to one another. Clicking on a Tweet in the decision-assistance panel scrolls the raw data stream to the active Tweet, additionally highlighting it in blue so that the subject can easily isolate it.



Social Media Analytics Interface Analysis Tool

Raw Data Feed - Live Twitter Stream sunglass hut

1. The Corinne Marie @TheCorinneMarie
San Diego Fashionistas! Join us Tues May 1st @sunglasshut Fashion Valley for the launch of @RadStyling!! 7-9 w/ @thepulsesd & @MrTopStyle4U

2. ki designs @pachuckidesigns
the #coach sunglasses aren't even listed on the #>sunglasshut website yet. eeek

3. Thats what she said! @easy_joke
RT @cubanmonkey69 Never new 'It was so hard' to merchandise a #sunglasshut kiosk! LMAO they have been @ it for 3 hrs lol..#TWSS

4. Liz Gallagher @cubanmonkey69
Never new it was so hard to merchandise a #sunglasshut kiosk! LMAO they have been @ it for 3 hrs lol

Decision-Assistance Panel (Top 20 Most Important Tweets)

2. ki designs @pachuckidesigns
the #coach sunglasses aren't even listed on the #sunglasshut website yet. eeek

3. Thats what she said! @easy_joke
RT @cubanmonkey69 Never new 'It was so hard' to merchandise a #sunglasshut kiosk! L
#TWSS

4. Liz Gallagher @cubanmonkey69
Never new it was so hard to merchandise a #sunglasshut kiosk! LMAO they have been @

38. King Shambles @FCShambles
@sunglasshut thanks for having an awesome return policy on new shades. #absolutequal

42. Natalie Mulford @XLaslife
Okay, so I will die if I can't get the PINK glitter Noir @miumiustlye sunnies. @sunglasshu
@luxottica_au

51. Dynafit @dynafitNA

Figure 2. Decision-assistance panel on right, raw data feed on left

4.1 Independent variable

We operationalized a single manipulated factor, social media analytics approach used in an organizational setting, as the contents of a decision-assistance panel intended to help participants analyze social media data. Panel contents were derived according to one of three approaches. Each treatment's panel comprised 20 Tweets for a balanced design.

The manual analysis condition received a random set of 20 continuous Tweets pulled from the raw data, instead of 20 individually random Tweets, as it is unlikely Tweets would be read out of order in a business setting. Understanding a given message might depend on reading it in series or embedded within a set of interactions. Thus, complete randomization of Tweets might unfairly suppress the control group's understanding by obscuring context that would be clear in a natural setting. To prevent biasing the manual treatment by inadvertently featuring a single random set of 20 consecutive Tweets featuring particularly useful or useless information, we also randomly assigned a different set to each subject.

The SA condition received the first of two decision-assistance panels compiled according to automated approaches. This panel consisted of 20 Tweets selected according to a word-based SA package in the R statistical application. The 10 most negatively and 10 most positively scored Tweets from the raw data were selected for inclusion, simulating the popular mechanism that drives a large percentage of currently-available free and fee-based social media monitoring systems.

Finally, the NLP condition received a panel of 20 Tweets selected through a natural language processing *simulation*, a necessary step as this technological capability is nascent and not yet robustly available. Due to the unavoidable dimension of humanness that cannot be separated from the process, we suggest that the output represents the effects of a "best NLP, " a superlative output benchmarking what we hope to eventually achieve with an NLP-based analytics system. Incremental analytical improvements can be compared to this best-case effort as machine algorithms steadily approach human capacity to understand unstructured text.

Operationalizations of all variables are specified in the following table.

Analytics approach to raw data interpretation	Treatment variable; expressed as contents of decision-assistance panel, 20 Tweets extracted from the raw data based on one of three methodologies. NLP group received 20 Tweets evaluated as most important by an NLP approach, SA group received top 10 most negative and top 10 most positive Tweets as evaluated by a sentiment analysis, and manual approach received random sets of 20 continuous Tweets.
Number	Self-reported counts of number of problems and number of opportunities identified.
Accuracy	Reflects external tally of problems and opportunities appearing both in subject's assessment and raw data. Standardized counts range from 1 to 7.
Confidence	Operationalized as a 1 to 5 Likert scale self-reported score.

Table 1. Operationalization of propositions and concepts in experiment

NLP Output Selection Procedure. To produce output approximately consistent with an NLP approach, we tasked three independent raters with manually executing a theoretical algorithm to pare 300 Tweets to the unequivocal top twenty most important customer-to-customer and customer-to-firm Tweets for SunglassHut management. The theoretical algorithm represents ideal processing steps that could be programmed into a machine to sift through unstructured social media texts, although such capability is not yet available in NLP software. Starting with 300 Tweets scraped directly from Twitter, sorters conducted the following steps:

I. Enhance signal-to-noise ratio – Eliminate advertisement messages not representing *unknown* or *undiscovered* intelligence; r**emove spam**; **tabulate retweets** (an amplification mechanism through which a Tweet is rebroadcast, essentially increasing signal strength and isolating potentially significant focal messages).

II. Refine signal – **Detect extreme sentiment** based on calculations of positive and negative word instances; d**etect suggestions** ("This product needs…"); d**etect requests** ("I need help with my…") that may serve as cues to identify important cues to undiscovered knowledge that could add value and potentially alter organizational decisions.

III. Signal Disambiguation – Resolve sarcasm potentially reversing polarity of apparent sentiment; **resolve anaphora** to determine which nouns back-referring phrases correspond with (Kao and Poteet 2007:1); **interpret slang, abbreviations, and paralinguistics**, or symbolic conventions used as shortcuts for standard concepts, phrases, or words (e.g., texting conventions and emoticons).

The NLP subfield of anaphora resolution is rich with algorithms demonstrating high rates of correct analyses (e.g., Kennedy and Boguraev 1996; Lappin and Leass 1994), while ambiguity clarification is the goal of many NLP tasks (e.g., parts-of-speech tagging (Hutchins 2006)) that depend on a range of types of human knowledge (grammatical rules, semantics, facts about the real world) for success (Mairesse et al. 2007). Ambiguity resolution requires "contextual sensitivity" and is extremely difficult to automate, in part because we lack accurately-labeled corpora for training machine learning systems (González-ibáñez & Wacholder 2011). Relevant to our goals, the use of hashtagged keywords by Twitter commentators to increase search accuracy has facilitated sarcasm corpus building as authors include "#sarcasm" to clarify intent. This is useful for contextualization.

IV. Socio-technical Calibration – Named entity extraction (NER) and **relationship extraction** provide additional clarification and may convey additional knowledge or alter meaning by identifying brands, businesses, particular people, etc. (Sang & De Meulder 2003) and disambiguate relationships between objects and people.

"Noise" Tweets immediately excludable by simple automated filters were eliminated, removing over half the raw data. Of the 140 remaining Tweets, 15 occurred on all three lists, 14 on two, and 46 on one. To test the hypothesis that the lists overlapped to such a degree merely due to chance, we consulted the hypergeometric distribution, which overlapping probability is known to follow (Fury et al. 2006). Based on $N = 140$, it is statistically significant that two lists of 30 overlap by 15 Tweets (p-value < 0.0001), so we proceeded on the assumption that the simulation outputs were due to the algorithm and not chance, particularly for the 15 Tweets occurring in all three outputs. For the remaining 5 Tweets, four new raters chose the 8 "most important" Tweets from the 14 appearing in two of the first three raters' outputs, based on the criterion of likely importance to Sunglass Hut management. Of these, 3 were selected by all raters, while 2 were chosen by three of the four; these 5 completed the decision-assistance panel for the NLP treatment. Thus, the NLP decision-assistance panel was rigorously designated according to input of seven independent raters.

4.2 Dependent variables

We are ultimately interested in how different social media analytics approaches support organizational decision-making. As proxy for this downstream construct, we measured the ability of participants to identify important problems and opportunities for the firm. We justify this operationalization by referring to the general assumption that organizational decision-making depends on information (Delbecq & Ven 1971; Galbraith 1974; Huber & McDaniel 1986) and that external problem and opportunity assessments are classic concerns of strategic planning (Houben et al. 1999). Opportunities convey chances to improve performance while problems are elements that could cause trouble for the business and therefore concern organizational managers.

Six dependent variables were measured, gauging the number of problems and opportunities identified by each participant from customer-to-firm or customer-to-customer Tweets (i.e., the raw data stream), the accuracy of participants' problem and opportunity assessments, and participants' confidence that they detected the important problems and opportunities for the firm from the raw data stream.

4.3 Control variables

Task type and raw data content were held constant by giving all participants the same task, objectives, instructions, and raw data from which to work. We also measured GPA and gender to test for possible differences in responses due to these influences (finding none).

5 Data Analysis

Eighty-five undergraduate MIS majors, ages 18 – 22, were recruited from a southeastern U.S. university campus and randomly assigned to one of three conditions, with 29 students in treatment 1, 29 in treatment 2, and 27 in treatment 3. Thirty reported as female (approximately 34%). MIS majors are considered appropriate participants in this social media-oriented study because they are likely candidates to intern in organizations that have implemented or are interested in implementing some type of social media analytics system. It is reasonable to expect a student intern to be assigned to manage or monitor this type of technology and report key information to managers for further analysis or decision-making. The students appeared to be engaged in the assignment and generally interested in the topic of research. Participants volunteered afterwards that the task was "fun."

A one-way MANOVA reveals a significant multivariate main effect for analytics approach, Wilkes' = 0.5236, p = 0.0001. Given the significance of the overall test, the univariate main effects are examined. Significant univariate main effects for analytics approach are indicated for **number** of problems identified, p = 0.00095, **accuracy** of problem assessment, p < 0.0001, **confidence** in problem detection, p = 0.0491, and **accuracy** of opportunity assessment, p = 0.0081. Due to unequal variances across treatment group responses for accuracy of opportunity assessment, we use the Kruskal-Wallis nonparametric test on this variable (Neter, Wasserman, and Kutner 1990: 642).

Significant treatment pairwise differences are obtained in a linear contrast of **number** of problems identified between NLP and SA, and NLP and random. The mean number of problems identified is 3.71 using SA, 5.86 using NLP, and 4.69 with a random set of Tweets. A similar pattern of pairwise differences is obtained for **accuracy** of problem assessment between NLP and SA, and NLP and random. The mean accuracy rating of problem assessments is 2.41 using SA, 4.38 using NLP, and 2.88 with a random set of Tweets. Finally, significant differences are obtained for **confidence** in problem detection between NLP and SA. The mean number of confidence levels indicated by participants is 3.1 using SA, 3.66 using NLP, and 3.19 with a random set of Tweets.

Significant treatment pairwise differences are also obtained for **accuracy** of opportunity assessment between NLP and SA, and NLP and random. The mean accuracy rating of opportunity assessments is 2.1 using SA, 3.17 using NLP, and 2.42 with a random set of Tweets.

ANOVA statistics for all dependent variables are presented in Table 2, including means of all measures.

	SA	NLP	Manual	Grand	Pr>F
Number – problems detected	3.71^{a}	**5.86**^{b}	4.69^{a}	4.77	0.000495*
Accuracy – problem detection	2.41^{a}	**4.38**^{b}	2.88^{a}	3.24	<0.0001*
Confidence – problem assessment	3.1^{a}	**3.66**^{b}	3.19^{a}	3.23	0.0491*
Number – opportunities detected	6.14	6.24	7.72	6.66	0.123
Accuracy – opportunity detection	2.1^{a}	**3.17**^{b}	2.42^{a}	2.57	0.008076*
Confidence – opportunity assessment	3.59	3.55	3.65	3.60	0.888

- Highest mean for each variable bolded and highlighted. Different superscripts indicate significantly different means for the four variables (bolded in first column) with significant ANOVA F-tests.

Table 2. Means and ANOVA results for dependent variables

Effect on Problem Number, Accuracy, and Confidence. Our investigation indicates significant variation in the ***number*** of problems identified by subjects across treatments. NLP enabled participants to identify a greater number of problems than either SA or manual approaches, though no difference is detected between SA and manual. **P4a** is supported while **P1a** is not.

Similar variation appears in the ***accuracy*** with which subjects identified problems from the data. The NLP approach enabled participants to more accurately assess problems than the other approaches, though again we detect no difference between SA and manual. **P5a** is supported while **P2a** is not.

Finally, the analytics approach introduces significant variation among participants' confidence in their ability to identify critical problems. Difference is found between NLP and sentiment-based approaches only. Thus **P6a** is supported by the data while P**3a** is not.

Effect on Opportunity Assessment Accuracy. The analytics approach to decision assistance also affects the ability of participants to accurately identify and assess opportunities from social media data. Our investigation indicates NLP enables a statistically greater degree of accuracy in identifying opportunities compared to SA and manual approaches, though no difference bears out between SA and manual approaches. **P5b** is supported while **P2a** is not.

Effect on Other Opportunity Variables. In contrast with the significant effect of the treatment on all problem detection measures, opportunity detection is not demonstrably affected along two of the three dimensions (number identified and confidence in assessment); these p-values are 0.123 and 0.888 respectively while power for these tests are 0.48 and 0.51. It is possible that "opportunity" is a fuzzier concept for students to grasp at a firm level, while "problems" are likely more straightforward to recognize. It may also lead to confusion that problems often can be restated as opportunities. For example, "lack of promotional pricing leads customers to defect to other brands" could be reformulated as the opportunity to provide more promotions in order to retain customer loyalty). However, the converse is not true—an existing opportunity is not likely reframable as an existing problem. Examination of the participants' opportunity assessments supports this notion. We noted problems reworded into opportunities such as "better customer service could increase customer base" and "educate the retailers on how to better merchandise the product."

Hypothesis	Supported?	Hypothesis	Supported?
1a: SA > manual (number of probs detected)	No	**4a:** NLP > SA and manual (number probs detected)	Yes
1b: SA > manual (number of opps detected)	No	**4b:** NLP > SA and manual (number opps detected)	No
2a: SA > manual (accuracy of prob detection)	No	**5a:** NLP > SA and manual (accuracy of prob detection)	Yes
2b: SA > manual (accuracy of opp detection)	No	**5b:** NLP > SA and manual (accuracy of opp detection)	Yes
3a: SA > manual (confidence in prob assessment)	No	**6a:** NLP > SA and manual (confidence in prob assessment)	Yes
3b: SA > manual (confidence in opp assessment)	No	**6b:** NLP > SA and manual (confidence in opp assessment)	No

Table 3. Summary of hypothesis testing

6 Results and Discussion

The most interesting result of the analysis is lack of support for our general conjecture that while only a very crude heuristic, sentiment analysis would still provide some measureable advantage beyond simply manually sifting through raw data. However, the results indicate that SA provides no advantage over simply reading random sets of consumer chatter. At the very least, these results should warn organizations to be cautious when attempting to link sentiment to actual business outcomes.

It is worthwhile to note that although participants did not identify greater numbers of opportunities or feel more confident about their opportunity identification using one analytics approach over another, accuracy of opportunity identification is still superior with the use of an NLP-based decision-assistance panel. It could be the case that number and confidence do not matter as much to ultimate organizational decision-making as accuracy; while the weighted importance of these factors is outside the scope of this research, this question nonetheless provides an interesting issue to address in future research.

It is also worthwhile to point out that for every dependent variable, manual analysis enables a ***slightly better*** performance that SA. Although not statistically significant taken singly, the probability that all six tests as a whole would favor manual analysis is 0.0156 according to the binomial distribution. In aggregation, SA is turns out to be worse than reading Tweets manually. Reliance on SA is thus likely detrimental to an analyst's ability to glean important information from social media data.

6.1 Limitations and future research

In order to benchmark a "best possible" NLP output, human sorters—subject to differences in opinions, interpretations, experience, and other factors not relevant to machines—executed the simulation. While we maintained rigor in our methods and gave full attention to precluding bias, we cannot escape the fact that the simulation was ultimately subject to human predisposition. As NLP capabilities become realized to a greater degree, further research replacing human simulation of machine algorithms with actual machine algorithms will be necessary to confirm whether this study's findings hold true to a truly automated NLP output. However, is important to have human analysis as a benchmark against which to compare future NLP algorithms, which this study provides. As we begin to implement these types of systems in practice, we will be better able to develop theory regarding how and why they work better or worse than systems based on other approaches, which will in turn allow us to improve upon system design and implementation.

While we have suggested potential explanations as to what factors might have confounded some of the opportunity measures, we do not know if there truly is no effect on opportunity number or confidence. Further, the sample of 85 participants is limited in size, resulting in an average cell size of 28. This potentially entails the limitations that accompany small sample sizes. We may lack power to detect a more subtle effect than that demonstrated on problem detection. In short, this study suffers from the general limitations associated with experimental research in a laboratory setting, implying that any generalizations of the findings must be applied with appropriate caution.

Additionally, we defer to those who point out that students are not superlative proxies for organizational decision-makers due to their limited context, background, and work experiences. In this study, however, we are not actually testing decision-making, but the ability to leverage social media text for eventual organizational decision-making that would be done by managers higher up in the organization. As such, we think students are appropriate subjects *in this study*. We argue that college-age adults are likely far better interpreters of what is being said in streams of social media text than executives not practiced in casual social media exchange on a daily basis.

Further, we are concerned with *differences* among the treatments, and less so with the absolute performances within a given treatment. While we agree that students may lack the experience and context for making certain decisions, this holds true across all of our treatments, so we believe bias would be systematic in this case, not obstructing our ability to detect differences due to treatments.

6.2 Conclusions

Few studies in Information Systems research take advantage of knowledge accumulated in qualitative text mining-related fields. However, the advent of the social media age commands attention to these technologies. Improved unstructured mining capabilities are likely to convey advantages to organizations willing to embrace novel approaches to better understanding their environment. Using advanced contextual mining methods to tap into the wealth of knowledge underlying customer-to-customer social-media-enabled exchange is forward looking and sophisticated, particularly compared to existing methods. We see exciting opportunities at this intersection of computer science, linguistics, organizational science, and IS.

This study is motivated by the presumption of a more useful objective than merely monitoring positive versus negative sentiment. While understanding customer sentiment is appropriate and relevant to a variety of research questions and consumer-oriented practices (e.g., peer reviews of products or movies), we suggest that the same capability can be subsumed much more accurately by an NLP-based mechanism, particularly as it applies to highly unstructured text. A comprehensive capability will enable social media analysts to detect sentimental extremes and, more importantly, discover a wide range of intelligence underlying customer comments, suggestions, requests for assistance, product-related issues, and other components that may prove important to decision-making despite being neither extremely positive nor extremely negative.

As the first experiment of a research program focused on the applicability of NLP to social media data, we suggest that future research can be formulated with reference to this study. The next step is to replicate this study with the incorporation of a fourth treatment, a decision-assistance panel derived by a machine-learning algorithm. This will allow us to document the disparity between "best possible" NLP and the state of the art of automated sorting by machine.

The most important contribution this study makes to social media research is to demonstrate that using sentiment analysis to learn from customers is likely less effective than humans reading streams of consumer chatter. This result is invaluable for improving social media monitoring practices. Empirical proof that an NLP approach is potentially superior to SA suggests that efforts to build an information system based on NLP techniques are a worthwhile and beneficial goal. NLP-based software promises the potential to substantially increase the knowledge firms may glean from tapping into customer-to-customer exchanges and enhance the effectiveness with which they monitor and respond to customer-to-firm communications.

7 References

Agichtein, E., Castillo, C., Donato, D., Gionis, A., & Mishne, G. (2008). Finding High-Quality Content in Social Media. Proc. of the Intl Conf on Web search and Web data mining, pp. 183–193, ACM.

Bunescu, R., & Mooney, R. (2007). Extracting Relations from Text: From Word Sequences to Dependency Paths (A. Kao & S.R. Poteet Eds.), Natural Language Processing and Text Mining (pp. 29–44). Springer.

Culnan, M. J., Mchugh, P. J., & Zubillaga, J. I. (2010). How Large U.S. Companies Can Use Twitter and Other Social Media to Gain Business Value. MIS Quarterly Executive, 9(4), 243–260.

Delbecq, A. L., & Ven, A. H. Van de. (1971). A Group Process Model for Problem Identification and Program Planning. The Journal of Applied Behavioral Science, 7(4), 466–492.

Demetriou, G., & Kawalek, P. (2010). Benefit-Driven Participation in Open Organizational Social Media Platforms: The Case of the SAP Community Network. Issues in Information Systems, XI(1), 601–611.

Eppler, M. J., & Mengis, J. (2004). The Concept of Information Overload: A Review of Literature from Organization Science, Accounting, Marketing, MIS, and Related Disciplines. The Information Society, 20(5), 325–344.

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text mining infrastructure in R. Journal of Statistical Software, 25(5), 1–54.

Fox, B. A. (1986). Local patterns and general principles in cognitive processes: Anaphora in written and conversational English. Text, 6(1), 25–51.

FreshNetworks. (2011). Social media monitoring report (p. 23).

Fury, W., Batliwalla, F., Gregersen, P. K., & Li, W. (2006). Overlapping probabilities of top ranking gene lists, hypergeometric distribution, and stringency of gene selection criterion. Proceedings of the

Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society, 1(m), 5531–5534.

Galbraith, J. R. (1974). Organization Design: An Information Processing View. Interfaces, 4(3), 28–36.

González-ibáñez, R., & Wacholder, N. (2011). Identifying Sarcasm in Twitter: A Closer Look. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers, 581–586.

Hearst, M. A. (1999). Untangling Text Data Mining. In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics, pp. 3–10, College Park, MD: Association for Computational Linguistics.

Hoffman, D. L., & Fodor, M. (2010). Can You Measure the ROI of Your Social Media Marketing? MIT Sloan Management Review, 52(1), 41-49.

Hotho, A., Andreas, N., & Paaß, G. (2005). A Brief Survey of Text Mining. Machine Learning, 20(1), 19–62.

Houben, G., Lenie, K., & Vanhoof, K. (1999). A Knowledge-based SWOT-analysis System as an Instrument for Strategic Planning in Small and Medium Sized Enterprises. Decision Support Systems, 26(2), 125–135.

Huber, G. P., & McDaniel, R. R. (1986). The Decision-Making Paradigm of Organizational Design. Management Science, 32(5), 572–589.

Hutchins, J. (2006). Example-based machine translation: a review and commentary. Machine Translation, 19(3-4), 197–211.

Kao, A., & Poteet, S. R. (2007). Natural Language Processing and Text Mining, (p. 272). Springer.

Kennedy, C., & Boguraev, B. (1996). Anaphora for Everyone: Pronominal Anaphora Resolution without a Parser. Proceedings of the 16th conference on Computational linguistics, pp. 113–118.

Lappin, S., & Leass, H. J. (1994). An Algorithm for Pronominal Anaphora Resolution. Computational Linguistics, 20(4), 535–561.

Mairesse, F., Walker, M.A., Mehl, M.R., & Moore, R.K. (2007). Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text. Journal of Artificial Intelligence Research, 30, 457–500.

Neter, J., Wasserman, W., & Kutner, M. H. (1990). Applied Linear Statistical Models (3rd Ed.), p. 1181.

Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval, 2(1-2), 1–135.

Pavlou, P. A., & Dimoka, A. (2006). The Nature and Role of Feedback Text Comments in Online Marketplaces: Implications for Trust Building, Price Premiums, and Seller Differentiation. Information Systems Research, 17(4), 392–414.

Rajman, M., & Besançon, R. (1998). Text Mining: Natural Language techniques and Text Mining applications. In S. Spaccapietra & F. Maryanski (Eds.), Data Mining and Reverse Engineering, pp. 50–64, Springer US.

Read, J. (2005). Using Emoticons to reduce Dependency in Machine Learning Techniques for Sentiment Classification. In Proceedings of the ACL Student Research Workshop, pp. 43–48, Ann Arbor, Michigan: Association for Computational Linguistics.

Sang, E. F. T. K., & De Meulder, F. (2003). Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL, pp. 142–147, Association for Computational Linguistics.

Speier, C., Valacich, J. S., & Vessey, I. (1999). The influence of task interruption on individual decision making: An information overload perspective. Decision Sciences, 30(2), 337–360.

Tanabe, L., Scherf, U., Smith, L. H., Lee, J. K., Hunter, L., & Weinstein, J. N. (1999). MedMiner: an Internet text-mining tool for biomedical information, with application to gene expression profiling. BioTechniques, 27(6), 1210–4, 1216–7.