https://doi.org/10.48009/1_iis_2014_430-433

CONSUMER SENTIMENT EXTRACTION FROM UNSTRUCTURED DATA

Matthew North, The College of Idaho, mnorth@collegeofidaho.edu
Samantha Riniker, The College of Idaho, samantha.riniker@yotes.collegeofidaho.edu

ABSTRACT

This paper represents the current state of an ongoing work in progress on the topic of customer sentiment analysis in unstructured data. The researchers have developed a body of classified customer postings from a variety of well-known companies, created a sentiment analysis matrix from those classifications, and are now at the point of validating a text mining model of their own design based on tested text mining algorithms. The process of developing these tools is described, along with some visual representations of the data and model being used. Once the proposed model is validated, likely after some fine tuning, multiple experiments will be carried out to prove the commercial applications and effectiveness of the model. The researchers intend to expand this work in progress into a paper for IIS 2015.

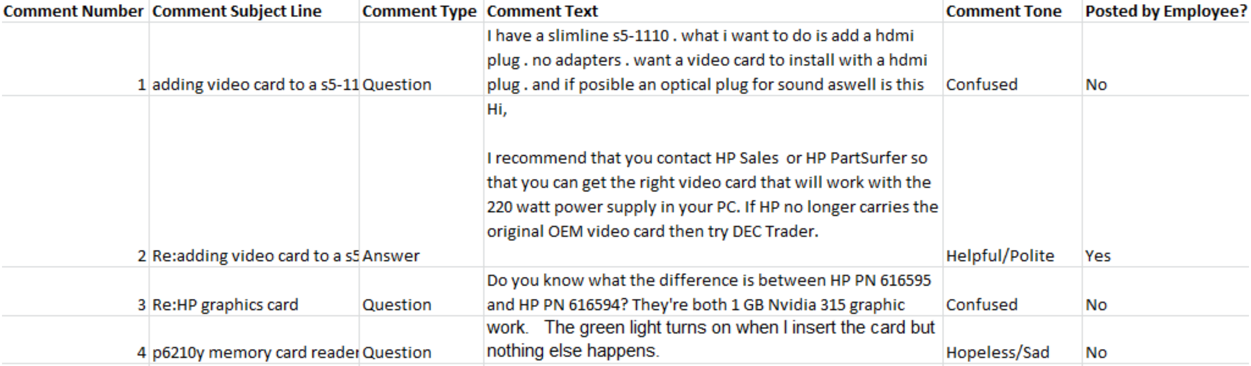
Keywords: Sentiment Analysis, Text Mining, Unstructured Data.

INTRODUCTION

Though estimates vary, industry experts agree that somewhere between 80 and 90 percent of data are unstructured in nature [1, 3, 7, 13]. Consumer discussion and help forums are one of many forms of unstructured business data, and the ability to quickly and accurately assess customer sentiment expressed online can be a powerful business analytics tool [10]. Companies that effectively monitor the pulse of customers can more quickly react to emerging issues, resulting in a higher rate of customer satisfaction and ultimately contributing positively to their own bottom line [2, 5]. This paper describes a work in progress wherein the authors are collaborating with a large group of undergraduate students to capture and categorize corporate discussion board postings into one of four sentiment categories. Categorized postings will be used as a training data set to validate a text mining model of our own design. With a validated model, we hope to then implement a functional sentiment extraction engine. Our plan is to develop this project into a full research paper to be published in *Issues in Information Systems* in 2015.

Research in Progress

We have developed a text mining model based upon multiple validated text mining algorithms for tokenization, stemming, gram generation, etc [6]. As of this writing, we are in the process of validating our model using the current corpus of categorized postings, presently with 7,250 observations in the data set (Figure 1).



Comment Number	Comment Subject Line	Comment Type	Comment Text	Comment Tone	Posted by Employee?
1	adding video card to a s5-11	Question	I have a slimline s5-1110 . what i want to do is add a hdmi plug . no adapters . want a video card to install with a hdmi plug . and if posible an optical plug for sound aswell is this	Confused	No
2	Re:adding video card to a s5	Answer	Hi, I recommend that you contact HP Sales or HP PartSurfer so that you can get the right video card that will work with the 220 watt power supply in your PC. If HP no longer carries the original OEM video card then try DEC Trader.	Helpful/Polite	Yes
3	Re:HP graphics card	Question	Do you know what the difference is between HP PN 616595 and HP PN 616594? They're both 1 GB Nvidia 315 graphic	Confused	No
4	p6210y memory card reader	Question	work. The green light turns on when I insert the card but nothing else happens.	Hopeless/Sad	No

Figure 1. Extract of categorized forum posts.

Once validated, we will run multiple experiments against uncategorized postings using k-Means Clustering, Discriminant Analysis and Decision Trees in an effort to extract consumer sentiment from customer discussion and help boards in real-time. In order to ensure model reliability, we are tracking data sources so that we can enforce experimental integrity during the second phase of the project. In the fall of 2014, we plan to present to IACIS conference attendees our categorization process, describe our data set, outline our sentiment model, and report progress on model validation and hopefully, share results of our first experiments using the model.

LITERATURE REVIEW

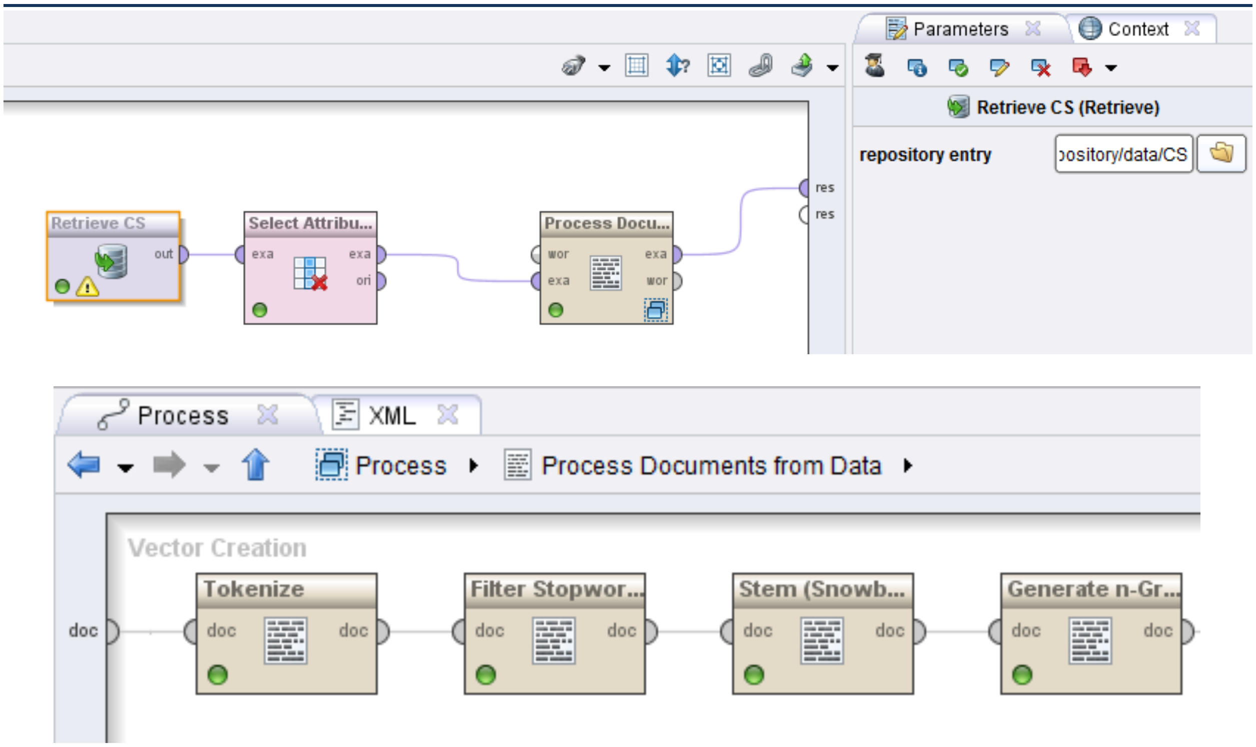
Text mining and the identification of customer sentiment from a range of sources is not new. Numerous researchers have engaged in studies to attempt to identify customer feelings and emotions in an effort to help companies better understand and serve their customers [7, 8, 13]. Research from Li, et al. is characteristic of both the questions and the approaches to using analytics to better understand customers [10]. Bush's brief article [4] illustrates the very nuts-and-bolts motivation for sentiment analysis, that is: Why do consumers behave the way they do? Ashbacher, et al., also investigate the ability to effectively identify customer emotions via automated systems [2].

Relevant to our study, Liao [11] and Mostafa [12] both examine customers' online line postings as unstructured data sources for understanding customer feelings. These studies establish our use of Snowball for word stemming, among other algorithms, as well as our planned use of Discriminant Analysis and k-Means algorithms to test our engine once our model is validated [also 6]. Zhan, et al. also use customer online comments as an unstructured data set, employing many of the same algorithms as well [14]. While additional research can be performed and other works cited, we are confident at this stage of our work in progress that we are well supported by published authorities in the fields of text mining and sentiment analysis.

BASIS OF STUDY

Beginning in November of 2013, some 130 undergraduate students began scouring the Internet for discussion board and help forums hosted by technology, retail, finance and health care companies. Companies represented include Apple, Google, Amazon, HP, Target, Highmark, and JP Morgan/Chase, among many others. Students captured and analyzed postings to topics within these forums, adding each to a central database. Data captured includes subject line, posting text, posting type, and posting source (e.g. customer, employee, community helper). Students then assign one of four sentiment category labels to each posting record: "Polite/Helpful", "Frustrated/Angry", "Confused", or "Hopeless/Sad" (see Figure 1). The researchers have text mined the categorized records for words which both appear frequently and carry emotional connotation capable of indicating consumer sentiment. Words bearing emotional connotation have been identified and validated in prior research, upon which we base our list [8, 14]. These words are then mapped to our four sentiment categories, forming a sentiment matrix against which our model can be tested.

Once our model has been tested against our sentiment matrix, we will evaluate the effectiveness of our model in terms of ability to map words in categorized postings to their assigned categories.



repository entry ɔository/data/CS

Figure 2. The Process/Subprocess comprising the present text mining model for this study.

High correlation of model category to assigned category will indicate a validated model. Moderate to low correlation will indicate one of several possibilities: 1) an invalid text mining model; 2) an unacceptably high rate of miscategorization of postings by student participants; 3) inaccurate word-to-sentiment mappings in our sentiment matrix. While we are hopeful our model will quickly yield high categorization correlation, we anticipate that will spend a portion of the summer of 2014 addressing some measure of the three items listed above, and fine tuning our model. Once we are satisfied that we have achieved a valid model, we will move to our second phase, focusing on running large extractions of discussion and help forum postings through our sentiment extraction engine. We anticipate testing by industry (technology, retail, finance, etc.), as well as randomly in order to determine the engine's effectiveness under a variety of circumstances. This methodology is common in our examination of previous research [9, 10, 12]. We expect to model our experiments after those which have yielded successful results in the past, but also the engage in experimentation of our own design anticipating not only useful results by following others' leads, but also by implementing our own novel text mining model in unique examinations of online postings. Should any such experiments be conducted prior to IACIS 2014, results will be shared and discussed with conference participants.

Implications

Our intent in this research is to build a viable and functional text mining model which would be immediately applicable to real-world business operations. Companies that use discussion and help forums would be able to adopt and use our engine to monitor and react to customer sentiment as it ebbs and flows in reaction to events which affect the business and its consumers. We believe that successful outcomes from our current work will add to the body of knowledge in the field of text mining and sentiment analysis/extraction, including to those works cited in this paper.

CONCLUSIONS

While much work remains to be done in this project, we are pleased with the progress we have made thus far. We believe that we have a suitable corpus of categorized forum postings and a model ready to be validated. Consumer sentiment extraction is at present an area of keen interest to almost any business, and the automation of such extraction represents significant possibilities for those companies to better understand, and respond to their customers. Based on existing literature, the tested algorithms underlying our model and the sentiment matrix we have developed, we anticipate that our present research will yield valid, reproducible, and eventually commercially useful analytic tool for unstructured business data.

REFERENCES

1. Amrich, D. (2013). Within Two Years, 80% of All Medical Data will be Unstructured. *zdNet, Electronic Version.* Retrieved on 27 March 2014 from: http://www.zdnet.com/within-two-years-80-percent-of-medical-data-will-be-unstructured-7000013707/
2. Aschbacher, H., Neukart, F., Kammerhofer, B., & Schatzl, S. (2009). The Use of Business Intelligence and Data Mining for the Detection of Customer Needs in Service Engineering. *Scientific Bulletin Series C: Fascicle Mechanics, Tribology, Machine Manufacturing Technology, 23*(100), 27-35.
3. Bridgwater, A. (2010). IBM: 80 Percent of Our Global Data is Unstructured (so what do we do?). *Computer Weekly, Electronic Version.* Retrieved on 27 March 2014 from: http://www.computerweekly.com/blogs/cwdn/2010/10/ibm-80-percent-of-data-is-unstructured-so-what-do-we-do.html
4. Bush, M. (2009). Text Mining Provides Marketers with the 'Why' Behind Demand. *Advertising Age, 80*(26), 14.
5. Chen, C.K., Shie, A.J., & Yu, C.H. (2012). A Customer-oriented Organisational Diagnostic Model Based on Data Mining of Customer-complaint Databases. *Expert Systems with Applications, 39*(1), 786-792.
6. Consoli, D. (2009). Analysing Customer Opinions with Text Mining Algorithms. *Computational Methods in Science and Engineering, 148*(2), 857-860.
7. Gantz, J. & Reinsel, D. (2011). Extracting Value from Chaos. EMC, Electronic Version. Retrieved on 27 March 2014 from: http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf
8. Iwashita, M., Shimogawa, S., & Nishimatsu, K. (2011). Semantic Analysis and Classification Method for Customer Enquiries in Telecommunication Services. *Engineering Applications of Artificial Intelligence, 24*(8), 1521-1531.
9. Jabr, W., Mookerjee, R., Yong T., & Mookerjee, V.S. (2014). Leveraging Philanthropic Behavior for Customer Support: The case of user support forums. *MIS Quarterly, 38*(1), 187-216.
10. Li, S.T., Shue, L.Y., & Lee, S.F. (2006). Enabling Customer Relationship Management in ISP Services through Mining Usage Patterns. *Expert Systems with Applications, 30*(4), 621-632.
11. Liao, S.H., Chen, C.M., & Wu, C.H. (2008). Mining Customer Knowledge for Product Line and Brand Extension in Retailing. *Expert Systems with Applications, 34*(3), 1763-1776.
12. Mostafa, M.M. (2013). More than Words: Social networks' text mining for consumer brand sentiments. *Expert Systems with Applications, 40*(10), 4241-4251.
13. Preimesberger, C. (2013). Managing Massive Unstructured Data Troves: 10 Best Practices. *eWeek, Electronic Version.* Retrieved on 27 March 2014 from: http://www.eweek.com/storage/slideshows/managing-massive-unstructured-data-troves-10-best-practices/
14. Zhan, J., Loh, H.T., & Liu, Y. (2009). Gather Customer Concerns from Online Product Reviews – A text summarization approach. *Expert Systems with Applications, 36*(2), 2107-2115.