Reliability and perceived value of sentiment analysis for Twitter data

JARI JUSSILA[1,a], VILMA VUORI[2], JUSSI OKKONEN[3], NINA HELANDER[1]

1 Information Management and Logistics, Tampere University of Technology, Tampere, Finland

2 Department of Management, University of Vaasa, Vaasa, Finland

3 School of Information Sciences, University of Tampere, Tampere, Finland

[a]Corresponding author: jari.j.jussila@tut.fi

Abstract: Social media offers rich data sources for companies that want to understand how they are perceived by their stakeholders. Sentiment analysis over Twitter can produce information about people's feelings towards their brand, business, and directors (Saif et al., 2012). Based on this information companies can take actions to enhance their customer experiences and perceived brand value. This study investigates the reliability and perceived value of two sentiment analysis tools developed to understand Finnish language, in contrast to human evaluators. For this purpose, a dataset of tweets from a Finnish software company was collected. For evaluating reliability Krippendorff 's α (Krippendorff, 2007) is computed. Perceived value of the automatic and human evaluator classified sentiment is evaluated by interviewing the case company representatives. The results point out that the analysis carried out by the human evaluators was perceived more valuable by the company representatives than the automatic analysis, due to different granulation level of the analysis. Compared to the automatic analysis, the human evaluators were able to put the identified emotions from the tweets better into a context, which in turn diminished the potential misinterpretation of who was the target of the most negative tweets.

Keywords: **social media, sentiment analysis, emotions, perceived value, brand**

Introduction

With the advent of social media, people have become more eager to express and share their opinions on web about corporate and product brands (Jansen et al., 2009). In the marketing literature customers' opinions and emotions are receiving increasing attention. Many studies have chosen Twitter as a source for collecting data on customers' opinions about brands (e.g. Jansen et al., 2009). One reason for this is that sentiment analysis over Twitter offers organisations a fast and effective way to monitor people's' feelings towards their brand, business, and directors (Saif et al., 2012). Sentiment refers in this study to an individual's state of negative or positive feeling that spreads through social interaction, that has an object and that ultimately aims to some kind of action (Jalonen, 2014).

Several computational approaches have been proposed to automatically identify and extract subjective information from tweets (Bravo-Marquez et al., 2014). Some

of these approaches are applicable to any language, while others are language specific. This study investigates the reliability and perceived value of SentiStrenght (e.g. Thelwall et al., 2012, 2010) and Nemo Sentiment and Data Analyzer (Paavola and Jalonen, 2015), both tools developed to understand Finnish language. The purpose of this study is to evaluate the reliability and perceived value of these two automatic sentiment analysis tools in contrast to analysis made by human evaluators. For this purpose, a dataset of tweets from a Finnish software company was collected. For evaluating reliability Krippendorff 's α is computed, which measures the agreement among observers, coders, judges, raters, or measuring instruments (Krippendorff, 2007). In order to evaluate the perceived value of these analyses and tools, interviews of the case company representatives were carried out.

Theoretical background

Social media enables free expression of vast range of sentiments that customers experience when interacting with a company or its products, services or brand. When a company is e.g. tagged to an emotionally charged tweet, it may have a significant positive or negative effect on company's brand (e.g. Jalonen, 2014) and consequently performance (Luo et al., 2013). Company brand is more and more affected by the way it is on display in social media (Khim-Yong et al., 2013). The simultaneous advantage and disadvantage of social media is that it promotes visibility, and its uncontrollable nature may multiply the implications for a company. The consequences of actions in social media are unpredictable but may be highly visible when going viral. In business setting this derives a need for making decisions and actions during a limited time span, as especially negative incidents in social media concerning the company require quick actions in order to diminish the risk of losing value (Jalonen, 2014). Companies can control the incidents in their external environment only to a certain limit, and therefore their only option is to try to understand these (Stoffels, 1994) and adjust their own actions accordingly.

Strategic management literature promotes the idea that a company's every action should be based on conscious decisions and that these decisions should be grounded on fair understanding of the current situation. Competitive intelligence is a process that analyses data and provides a company relevant information about the external environment and thus helps the company gain competitive advantage over other players in the field (e.g. Bensoussan & Fleisher, 2007; Vuori, 2011). One of the most contemporary competitive intelligence actions aimed to back up decision making is analysing data derived from social media. In fact, social media offers a new kind of dimension to competitive intelligence - the social aspect of the masses, providing new kinds of analytics and a path to transform social media content into strategically actionable knowledge (He et al., 2016). For example, understanding how the company brand is perceived by customers and how their sentiments towards it are expressed in social media (Stieglitz & Dang-Xuan, 2013) gives the company possibility to take actions to enhance the customer experience and promote brand visibility.

Sentiment analysis software is seen as an efficient way to analyse the masses of social media derived data providing companies understanding of how it is depicted in social media. However, the reliability of sentiment analysis software may be questioned. Firstly, analysis made with sentiment analysis software is based on binary machine logic, where the data is refined by a system lacking serendipity and ambiguity. This may cause the risk of false analysis, as the machine logic may not correctly interpret the content of tweets made by human beings. Non-binary human logic may contain many different truths due to its fuzziness and ambiguity, whereas binary machine logic allows only one, possibly misinterpreted, truth (Vuori & Okkonen, 2012). Another noteworthy issue is that while the subject of information may be internal (e.g. the company brand), the sources of the information may be both internal (employees) and external (customers). The company may be tagged in a tweet by an employee, and it is likely that the employees' tweets are more often positively than negatively charged. It is worth of discussion, should these "insider tweets", possibly distorting the data, be extracted from analysis that focuses on customer experience and brand perception. Furthermore, automatic analysis tools are not usually able to take into account the context of the tweet, which may further reveal the target of the expressed emotion. And finally, the automatic analysis tools do not typically understand sarcasm, that e.g. tweets can include. Taking to account these frailties of using software to analyse sentiments, it is fair to ask can it in fact understand the sentiments expressed in social media, and, more importantly, Is it safe to base decisions of company's actions on such analysis?

Research approach

Computational approaches and tools that can understand Finnish language include SentiStrength (e.g. Thelwall et al., 2012, 2010) and Nemo Sentiment and Data Analyzer (Paavola and Jalonen, 2015). This study investigates the reliability and perceived value of these two automatic tools, in contrast to human evaluators. Nemo Sentiment and Data Analyzer tool is a cloud-based service that enables both collecting the Twitter data and analyzing sentiment using two separate algorithms: one based on logistic regression (LR) and the other on random forest (RF) classification. SentiStrenght algorithm calculates the positive and negative sentiment strength for each tweet on a scale of 1 to 5. These values were used to compute the classification of the tweet to positive, neutral or negative. The three algorithm based data classifications are compared to two human evaluator classifications. The working hypothesis is that human evaluators classify tweet data uniformly and are able to extract correct sentiments by human logic.

In order to investigate the reliability and perceived value of automated vs. human evaluator evaluated sentiment analysis, we collected Twitter data from a Finnish software company. A total of 509 tweets were collected using Nemo Sentiment and Data Analyzer tool.

The human evaluators independently classified the tweets into one of three categories: positive, neutral or negative using a spreadsheet processor. The data was

imported to SentiStrenght and sentiment strength was calculated for each tweet. Krippendorf's α (Hayes & Krippendorff 2007) value was then calculated using SPSS to the human evaluator classified tweets, SentiStrenght classified tweets, and Nemo Sentiment and Data Analyzer tool classified tweets. In the process the evaluations were compared in pairs assuming that human evaluations were the baseline, yet they had distinctive notions on the sentiments. For avoiding misunderstanding and misinterpretations α should have somewhat high value. Social scientists commonly rely on data with reliabilities α ≥ ,800, consider data with ,800 > α ≥ ,667 only to draw tentative conclusions, and discard data whose agreement measures α < ,667 (Krippendorff 2004). In this case all values of are low as depicted in Table 1.

For the evaluation of the perceived value of the analyses, we carried out a workshop where the results of the analysis were presented to the case company representatives (business unit manager, key account manager, marketing specialist) and their opinions on the analysis and its value for business decisions were asked.

Results

The evaluations are presented in Table 1. On each row evaluations are compared on absolute levels and by Krippendorff's α. The first three columns describe the values difference in assessment as the nominal difference between evaluator is described with bolded figure on each row.

Table 1. Evaluations.

	Positive agreement	Neutral agreement	Negative agreement	Krippendorff's α
Human evaluator 1 vs Human evaluator 2	337 vs 171	146 vs 326	19 vs 5	,1962
	166	**180**	**14**	
Human evaluator 1 vs Nemo LR	337 vs 121	146 vs 348	19 vs 20	-,0492
	216	**202**	**1**	
Human evaluator 1 vs Nemo RF	337 vs 224	146 vs 217	19 vs 13	,1787
	113	**71**	**6**	
Human evaluator 1 vs SentiStrenght	337 vs 265=	146 vs 227	19 vs 10	,2700
	72	**81**	**9**	
Human evaluator 2 vs Nemo LR	171 vs 121=	326 vs 348	5 vs 20	,3266
	50	**22**	**15**	
Human evaluator 2 vs Nemo RF	171 vs 224	326 vs 217	5 vs 13	,2760
	53	**109**	**7**	
Human evaluator 2 vs SentiStrength	171 vs 265	326 vs 227	5 vs 10	,3102
	94	**99**	**5**	
Nemo LR vs Nemo RF	121 vs 224=	348 vs 217	20 vs 13	,4563
	103	**131**	**7**	
Nemo LR vs SentiStrength	121 vs 265	348 vs 227	20 vs 10	,1821

	144	**121**	**10**	
Nemo RF vs SentiStrenght	224 vs 265	217 vs 227	13 vs 10	,2057
	41	**10**	**3**	

The interviewed company representatives perceived the analysis valuable in general terms, as they were able easily and in visual way to see the distribution of the negative vs. positive tweets hashtagged to their company. When taking a closer look to the tweets analysed by the human evaluators, the company representatives found the influence of the context highly relevant. For example, most of the negative tweets were not targeted towards the case company, but instead e.g. towards the Finnish government making unwise decisions concerning information systems, or even towards the case company's competitors. This important fact was not revealed by the automatic analysis tools, as they were not able to take into account the context of the specific tweet, nor the potential sarcasm behind the tweet.

Conclusions and discussion

As noted in results section, all pairs failed the test in sense that the hypothesis of human being more powerful and smart algorithms existing should be discarded. Drawn from the data there is significant variations with human and machine evaluations, yet algorithms provide more uniform analysis. There are several reasons for such variation. First reason for failure is the limited amount of data, there can be structural issue that hinder the tweet classification by the human evaluators. On the other hand the use of human evaluators is dependent on their subjective experience on the issue and personal attributes. The high variation in classification is due to human evaluators different judgement.

Highest values of α was when compared human evaluator 1 or human evaluator 2 to SentiStrenght or Nemo LR to Nemo RF. This provides initial evidence, that elaborated algorithms and human evaluators may succeed, yet it is issue of technical development. On the other hand, Nemo LR and Nemo RF provided highest α, due to common ancestry as they assess the tweets similarly.

Human evaluators provide assessment based on their insight and/or prior knowledge. However in this data the assumption of human sensitiveness or preciseness on sentiments is not supported due to high variation. The study does however, have several limitations that may impact the results. The human evaluators had little prior experience of performing sentiment analysis and their knowledge of the company and its business was limited. This can, for instance, impact the interpretation on what tweets are actually classified negative towards the company. In addition, the amount of observations was relatively small and not meeting gold standards set for sentiment classification. Further research should include a more extensive set of observations and involve also company employees and more experienced sentiment evaluators performing the classification.

References

Bensoussan, B. E., & Fleisher, C. S. (2007). Business and Competitive Analysis: Effective Application of New and Classic Methods. New Jersey: Financial Times Prentice Hall.

Bravo-Marquez, F., Mendoza, M., Poblete, B., (2014). Meta-level sentiment models for big social data analysis. Knowl.-Based Syst. Vol. 69, pp. 86–99.

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. Communication Methods and Measures, 1, 77-89.

He, W.,Tian, X., Chen, Y. & Chong, D. (2016) Actionable Social

Media Competitive Analytics For Understanding Customer Experiences, Journal of Computer

Information Systems, 56:2, 145-155

Jalonen, H. (2014). Negatiiviset tunteet ja sosiaalinen media muodostavat yrityksille vaikean yhdistelmän. LTA, 2, 14.

Jansen, B.J., Zhang, M., Sobel, K., Chowdury, A., (2009). Twitter power: Tweets as electronic word of mouth. J. Am. Soc. Inf. Sci. Technol. Vol. 60, pp. 2169–2188.

Khim-Yong, G., Cheng-Suang, H., & Zhijie, L. (2013). Social Media Brand Community and Consumer Behavior: Quantifying the Relative Impact of User- and Marketer-Generated Content. Information Systems Research, 24(1), 88-107.

Krippendorff, K. (2004). Reliability in Content Analysis: Some Common Misconceptions and Recommendations.Human Communication Research, 30(3), 411-433.http://dx.doi.org/10.1111/j.1468-2958.2004.tb00738.x

Krippendorff, K., (2007). Computing Krippendorff's alpha reliability. Dep. Pap. ASC 43.

Luo, X., Zhang, J., & Duan, W. (2013). Social media and firm equity value.Information Systems Research, 24(1), 146-163.

Paavola, J., Jalonen, H., (2015). An Approach to Detect and Analyze the Impact of Biased Information Sources in the Social Media, in: Proceedings of the 14th European Conference on Cyber Warfare and Security 2015: ECCWS 2015. Academic Conferences Limited, p. 213.

Saif, H., He, Y., Alani, H., (2012). Semantic sentiment analysis of twitter, in: The Semantic Web–ISWC 2012. Springer, pp. 508–524.

Stoffels, J. D. (1994). Strategic issues management: A comprehensive guide to environmental scanning. Pergamon.

Thelwall, M., Buckley, K., Paltoglou, G., (2012). Sentiment strength detection for the social web. J. Am. Soc. Inf. Sci. Technol. Vol. 63, pp. 163–173.

Thelwall, M., Buckley, K., Paltoglou, G., (2011). Sentiment in Twitter events. J. Am. Soc. Inf. Sci. Technol. Vol. 62, pp. 406–418.

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., Kappas, A., (2010). Sentiment strength detection in short informal text. J. Am. Soc. Inf. Sci. Technol. Vol. 61, pp. 2544–2558.

Stieglitz, S. & Dang-Xuan. L. (2013) Emotions and Information Diffusion in Social Media—Sentiment of Microblogs and Sharing Behavior. Journal of Management Information Systems. Vol. 29, Iss. 4, 2013

Vuori, V., & Okkonen, J. (2012). Refining information and knowledge by social media applications: Adding value by insight. Vine, 42(1), 117-128.

Vuori, V. (2011). Social media changing the competitive intelligence process: elicitation of employees' competitive knowledge. Tampereen teknillinen yliopisto. Julkaisu-Tampere University of Technology. Publication; 1001.