A Revised Lexical Approach for Analyzing Online Reviews

Xiaowen Fang[1] and Fan Zhao[2]

[1] School of Computing, DePaul University, Chicago, IL, U.S.
xfang@cdm.depaul.edu
[2] Lutgert College of Business, Florida Gulf Coast University, Fort Myers, FL, U.S.
fzhao@fgcu.edu

Abstract. Inspired by the lexical approach used by psychologists to study personality traits, this paper proposes a revised version of this approach for analyzing online reviews. The lexical approach is based on a lexical hypothesis stating that personality traits are reflected in the adjectives invented by people to describe them. The revised lexical approach contains five steps: collecting online reviews, parsing adjectives, extracting consumer/user observations, factor analysis, and exploring factors/patterns. The paper elaborates each of these steps. It further discusses implications of this new approach.

Keywords: lexical approach, content analysis, qualitative research, online reviews.

1 Introduction

Web 2.0 technologies such as blogs, wikis, and forums have created a perfect environment for content sharing. Consumers can share their opinions about a product or service with large audiences with minimal efforts (such as a few mouse clicks). These online reviews often contain critical information for both practitioners and researchers in an unprecedented scale. There have been numerous reports suggesting the increased use of online content for both academic research and business decision-making [7].

Content analysis, a class of methods at the intersection of the qualitative and quantitative traditions, has been commonly applied for rigorous exploration of many important but difficult-to-study issues of interest to management researchers [4]. When applying the content analysis method to online reviews, there are two major problems that may jeopardize the quality of the analysis results.

- The analysis of texts and interpretation of results are subjective. Researchers who specialize in content analysis research have expressed major concerns of this method about a disconnect between what the content analysis results can tell readers legitimately versus how the findings are actually being interpreted by the authors [2, 4]. This disconnect is caused by not only researchers' subjective interpretation of results, but also by extensive human coding of texts.

M. Kurosu (Ed.): Human-Computer Interaction, Part I, HCII 2014, LNCS 8510, pp. 69–76, 2014.

- The sheer volume of online reviews poses a major problem in content analysis. For example, Simmons et al. [7] report a computer aided content analysis of over 20,000 movie reviews. The number of online reviews can easily rise up to a million or more in some contexts. It is impractical to use human judges to code comments presented in online reviews.

Researchers have attempted to apply natural language processing technologies to improve the efficiency and effectiveness of content analysis [4, 7]. While to some extent, the efficiency of content analysis can be improved to deal with a large number of reviews, quality of the analysis can be compromised by the fact that most online reviews tend to be informal and do not adhere to accurate grammar rules. Any attempt to accurately parse semantics in online reviews will be proved counter-productive. It is argued that the real essence of content analysis is to explore consumers' language and discover its true meanings. Super-imposing formal grammar rules to a consumer language that one tries to investigate is no doubt very dangerous and may likely skew the results.

Inspired by the lexical approach adopted by psychologists to study personality traits [1], this paper proposes a revised lexical approach aiming at addressing the aforementioned limitations of current content analysis method: namely subjective judgments, large number of online reviews, and inappropriate use of natural language processing technologies. This revised lexical approach is promising and will likely contribute to both research and practice in any area concerning human behaviors such as human computer interaction, information systems, management, and psychology by substantially improving the rigor of content analysis of online reviews.

The rest of the paper is organized as follows: it first introduces the lexical approach used by psychologists to study human personality. After the following section proposes the lexical hypothesis, Section "A Revised Lexical Approach" elaborates details of the revised lexical approach for analyzing online reviews.

Then the paper presents discussions about the strengths and potential weaknesses of the revised lexical approach.

Finally, the last section discusses how this revised lexical approach can be applied in research and practice, and its profound implications.

2 The Lexical Approach Used in Personality Study

The idea of using a lexical approach to obtain personality traits stems from the lexical hypothesis for personality research. The lexical hypothesis states that people will want to talk about personality traits that they view as having important consequences in their lives [1]. As a result, people will inevitably invent some words to describe those who exhibit high or low levels of these essential traits. Over long periods of time, words that describe important traits should become established in every language [1]. In applying a lexical approach to personality research, a researcher first systematically searches the dictionary of the language to be examined in order to obtain a list of personality–descriptive adjectives [1]. After establishing this list of adjectives, the researcher excludes terms that are rarely used. The resulting list is then

administered to a large sample of participants who are asked to provide self-ratings on these adjectives, indicating the extent to which each adjective describes their own personalities. A factor analysis is then performed on the ratings collected through the survey. Each factor discovered in this factor analysis constitutes a unique personality trait. The list of words converging on a factor suggests the nature of this personality trait and how to describe this trait.

Cattell [3] was the first researcher to conduct a factor analysis of ratings on personality-descriptive adjectives in the English language. His analysis revealed 12 factors. However, when other researchers later re-analyzed Cattell's data, they found only a consistent set of five personality factors[8]. These five personality factors later became the well-received Big Five personality factors to describe personality traits – openness, conscientiousness, extraversion, agreeableness, and neuroticism [5].

This paper argues that the lexical approach used in personality research can be an invaluable method in content analysis of online reviews for two reasons: 1) this approach effectively reduces the unit of analysis to a single word without necessarily interpreting its meaning beforehand. 2) The factor analysis in this approach leads to the discovery of any significant patterns that may exist in the language used by reviewers. Overall, this approach provides a possible solution to avoid potential pitfalls of subjective judgments by researchers in content analysis while leading to the findings of patterns in reviewer language.

However, the lexical approach cannot be applied in content analysis of online reviews before the following three issues are addressed:

1. The lexical hypothesis must reasonably stand for the text content to be analyzed. The lexical approach hinges on this hypothesis. If it doesn't hold true, an attempt to apply the lexical approach will be futile.
2. The lexical approach requires the use of a list of adjectives that describe a research object such as a product or service. Unlike human personality, researchers usually do not have full knowledge about the language used by consumers about a product or service. No dictionary of adjectives is available for online reviews.
3. The survey conducted towards the end of the lexical approach may become an unconquerable barrier for many studies. In order to achieve reliable and robust results in the factor analysis, this survey may require thousands of responses. The survey is clearly too time-consuming and inefficient.

To move forward with the lexical approach, the next section addresses the essentiality of the lexical approach and Section "A Revised Lexical Approach" proposes a revised lexical approach applicable to content analysis of online reviews.

3 The Lexical Hypothesis

The lexical hypothesis states that a finite set of traits of a product or service are consistently perceived by its consumers and these traits can be expressed by consumer language. This hypothesis is the premise of applying the lexical approach in content analysis of online reviews. In general, this hypothesis likely holds true if an online

community focusing on a product or service where consumers can share their opinions have been formed for an extended period of time and the amount of online reviews is substantial. For example, lexical hypothesis may likely hold true for computer games because game players have posted reviews and formed online communities for years.

4 A Revised Lexical Approach

Attempting to address the issues associated with content analysis and the original lexical approach in personality research, this paper proposes a revised lexical approach. This approach has five major steps: 1) Collecting online reviews, 2) Parsing adjectives, 3) Extracting consumer/user observations, 4) Factor analysis, and 5) Exploring factors/patterns. Details about each step are discussed in the following subsections.

4.1 Collecting Online Reviews

The goal of this step is to gather user-generated content from the Internet. Most of online reviews are available for downloading on the Internet. To ensure a good quality of the downloaded reviews, one has to attend to the following details:

1) The downloaded reviews should be representative. The reviews must include views and opinions expressed by different stakeholders such as vendor/manufacturer, retailer, and users, and different product types.
2) Only the reviews posted by users will be included. Any texts appearing on a website as standard elements such as headers and footers should be excluded.
3) Repetitive content must be removed. Many users post their reviews in response to others'. The repetitive content may skew the results.
4) Reviews must be separated by both reviewer and product. Each review should have been posted by one reviewer about one product/service.
5) A database needs to be designed to store structural information such as product name, reviewer information, and time of the review.

Each review has to be stored as one record in a database. The meta-information about each review is vital for interpreting the results from factor analysis in a latter stage.

4.2 Parsing Adjectives

Since no known dictionary of adjectives likely exists for a regular consumer product or service, it is imperative to identify such adjectives from online reviews. A computer program that applies the most basic Natural Language Processing (NLP) techniques can be used/developed to perform the following tasks:

1) Parsing words. The NLP application first breaks texts into words. Any special characters or white spaces are removed.
2) Checking the part of speech. The NLP application then connects to a lexical database for the English language such as WordNet[6], to check the type of part of speech of each word.
3) Detecting adjectives that describe the target product or service. These adjectives will then be saved as the candidate words to be further analyzed.

In this step, it is critical NOT to apply any sophisticated NLP techniques attempting to shorten the list of adjectives or interpret semantic meanings of those words. The assumption is that consumers or users have their own language that is not necessarily the same as the standard English. Researchers ought to keep the adjectives used by consumers/users intact and free from external interference. If there are any prominent patterns in the consumer/user language such as phrases and synonyms, they will surface in later analyses.

4.3 Extracting Consumer/User Observations

As discussed in Section 2, one of the challenges to directly apply the lexical approach in the analysis of online reviews is the difficulty of conducting a large-scale survey. Fortunately, the online reviews themselves represent invaluable consumer/user observations that can be used to substitute the survey in the original lexical approach. Each online review can be treated as an individual observation made in a natural environment. Only the most important and most appropriate adjectives would have been chosen in these reviews. Consumers/users voluntarily contributed these reviews without any external pressure. Even though the online reviewers weren't given the opportunity to choose their words from the complete list of adjectives used by their peers, this paper argues that the online reviews constitute a huge dataset with higher quality than the responses from a hypothetical survey conducted in a later stage. The size of such a dataset is unprecedented and could have never been achieved in a regular survey. Arguable, selecting words from a long list can be tedious, time-consuming, and inaccurate. Ultimately, the quality of a hypothetical survey can be at risk due to the fatigue factor. The huge amount of online reviews also helps find patterns that may never be found through other means.

To prepare for the upcoming factor analysis, the online reviews must be converted to a dataset by a computer program as follows: 1) Each word on the list of adjectives produced in the second step, "Parsing Adjectives", is treated as an individual item. The list of adjectives can be saved as the field names of a database table. 2) Retrieve all online reviews one at a time. Each review about one product/service is treated an individual record. Adjectives used in the same review must be somehow related because they describe the same product/service. If an adjective appears in this review, the value for this adjective (field) shall be 1. Otherwise, a zero value should be registered.

The end product of this step is a table of values "0" and "1". "1" indicates the appearance of an adjective in a particular online review while "0" suggests absence of the word. This table is used as the dataset for the upcoming factor analysis.

4.4 Factor Analysis

An exploratory factor analysis is conducted to discover the patterns of adjectives used by consumers. Each of the factors surfaced in this analysis represents a small list of adjectives that share some commonalities because they have been used together for some reason. These commonalities may provide critical information to researchers regarding different perspectives of the target product or service. Although they are subject to further interpretation, there is no doubt that the existence of these commonalities is factual.

4.5 Exploring Factors/Patterns

The purpose of this step is to interpret the meanings and implications of each factor discovered in the previous factor analysis. Qualitative methods such as interviews can be used to help understand each factor. Although there is still a subjective component in this process, the patterns of adjectives are derived from actual data. These patterns are robust and can reasonably withstand biases from human judgments. These factors represent common issues or concerns expressed by the majority of consumers or users. The spectrum of these factors will likely cover all perspectives of a product/service. This information is invaluable to all stakeholders that are involved in the entire life cycle of a product/service.

5 Discussions

This paper proposes a revised lexical approach for analyzing online reviews. This promising approach will benefit both researchers and practitioners in any subject area that concerns about human behavior. The strengths of this revised lexical approach are summarized as follows:

- It significantly improves the rigor of content analysis of online reviews by introducing a well-established statistical analysis into a subjective and qualitative process of content analysis. The results from the factor analysis should clear any doubts about whether or not a pattern observed from the analysis is factual or not. The data speaks for itself.
- The proposed approach makes it possible to use an unprecedentedly large sample that has never been achieved before. Due to the benefits of a large sample, the thoroughness and validity of this analysis can reach a higher level than what researchers have ever attempted to achieve.
- Data collection is naturalistic and non-intrusive. Most online reviews posted by ordinary customers are typically written in a natural environment that is free of

external influence such as financial incentive, time pressure, biased instructions from the experimenter, fatigue, and the like. The quality of the collected data may be higher than that of the data collected in most empirical studies we have seen so far.

- The proposed approach is economic. The guiding principle of this approach is to let the data speak for itself. No interpretation of any word, sentence, or review written by consumers or users will be necessary. No sophisticated natural language processing techniques are ever used. All computer programs involved in the analysis can be programmed by a student who possess a Bachelor's degree in computer science or a graduate student majoring in any IT field.

However, as any research method in general, the revised lexical approach is not perfect either. There are two main constraints of this method:

- By analyzing only adjectives in online reviews, the revised lexical approach might have overlooked some useful information expressed through other parts of the languages such as nouns. Additional analyses of other parts of speech such as nouns can be conducted to complement the analysis of adjectives.
- The environment where an online review is written cannot be controlled. A reviewer doesn't have an opportunity to assess the applicability of all possible words from a complete dictionary in an online review. A score of 0 or 1 might have oversimplified a reviewer's evaluation of a word.

6 Conclusions

The revised lexical approach proposed in this paper may have profound implications in both academia and industry. It provides a rigorous quantitative solution to a problem that used to be solved only by subjective and qualitative methods. It is conceivable that this method can be applied in a wide spectrum of research disciplines.

References

1. Ashton, M.C.: Individual Differences and Personality. Academic Press, San Diego (2007)
2. Carlson, L.: Use, Misuse, and Abuse of Content Analysis for Research on the Consumer Interest. The Journal of Consumer Affairs 42(1), 100–105 (2008)
3. Cattell, R.B.: Confirmation and clarification of primary personality factors. Psychometrika 12, 197–220 (1947)
4. Duriau, V.J., Reger, R.K., Pfarrer, M.D.: A Content Analysis of the Content Analysis Literature in Organization studies-research themes, data sources, and methodological refinements. Organizational Research Methods 10(1), 5–34 (2007)
5. Goldberg, L.R.: An alternative 'description of personality': The Big-Five factor structure. Journal of Personality and Social Psychology 59(6), 1216–1229 (1990)

6. Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.J.: Introduction to WordNet: an on-line lexical database. International Journal of Lexicography 3(4), 235–244 (1990)
7. Simmons, L.L., Conlon, S., Mukhopadhyay, S., Yang, J.: A computer aided content analysis of online reviews. The Journal of Computer Information Systems 52(1), 43–55 (2011)
8. Tupes, E.C., Christal, R.E.: Recurrent personality factors based on trait ratings. Journal of Personality 60, 225–251 (1992)