Decision Support Systems 59 (2014) 262–273

Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

External validity of sentiment mining reports: Can current methods identify demographic biases, event biases, and manipulation of reviews?

Fons Wijnhoven *, Oscar Bloemen

University of Twente, Enschede, Netherlands

ARTICLE INFO

Article history:
Received 5 June 2013
Received in revised form 22 November 2013
Accepted 16 December 2013
Available online 25 December 2013

Keywords:
Sentiment mining
Opinion mining
External validity
Demographic bias
Event bias
Product review manipulation
Design proposition validation

ABSTRACT

Many publications in sentiment mining provide new techniques for improved accuracy in extracting features and corresponding sentiments in texts. For the external validity of these sentiment reports, i.e., the applicability of the results to target audiences, it is important to well analyze data of the context of user-generated content and their sample of authors. The literature lacks an analysis of external validity of sentiment mining reports and the sentiment mining field lacks an operationalization of external validity dimensions toward practically useful techniques. From a kernel theory, we identify multiple threats to sentiment mining external validity and study three of them empirically 1) a mismatch in demographics of the reviewers sample, 2) bias due to reviewers' incidental experiences, and 3) manipulation of reviews. The value of external validity threat identifying techniques is next examined in cases from Goodread.com. We conclude that demographic biases can be well detected by current techniques, although we have doubts regarding stylometric techniques for this purpose. We demonstrate the usefulness of event and manipulation bias detection techniques in our cases, but this result needs further replications in more complex and more competitive contexts. Finally, for increasing the decisional usefulness of sentiment mining reports, they should be accompanied by external validity reports and software and service providers in this field should incorporate these in their offerings.

1. Introduction

Knowledge of the experiences clients have with a company and their competitors' products and services is crucial, as responding correctly to this information can lead to a competitive advantage [1]. Nowadays, acquiring this knowledge can be assisted by using the ever growing number of sentiments-containing expressions publicly available in (micro-)blogs, review sites, and forums [2]. Manual examination of all these data is a daunting task and automation is desirable.

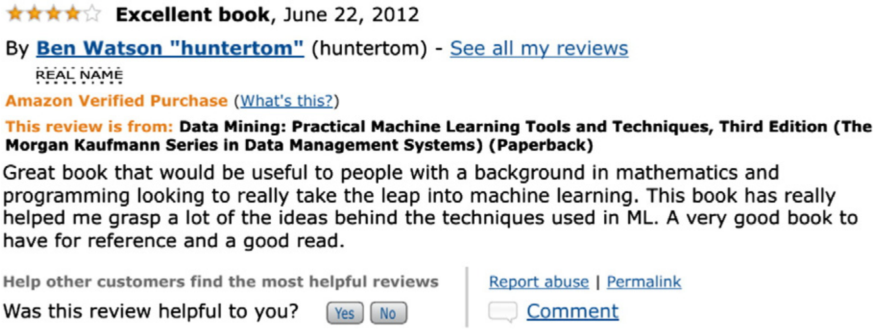
Solutions for the automated extraction of sentiments come from a subsidiary of machine learning, named opinion or sentiment mining [3,4]. Determining the sentiment of a text is in essence a classification problem with classes positive, negative [2] and neutral [5]. Given an opinionated text as in the book review of Fig. 1, a classifier may determine the polarity of the sentiment by comparing words in the text with words in a lexicon of which the polarity is known. In the case of this book review, words like "great", "helped", and "good" indicate a positive review. Analyzing this text with sentiment mining tool Pattern [6] shows that the review is 0.19 positive on a scale from −1 to 1.

The various applications of sentiment mining span a large domain, like movies [4] commercial products and services [3,7,8], product features [9,10] also on a comparative basis [11], and the sentiment toward a political party or topic [2]. In the majority of sentiment mining research, the dominant topic is the classification algorithm. The algorithms are continuously improved to squeeze out the last percentage increase in accuracy [12]. However, if the goal of sentiment mining is harvesting market or public information for decision making, it is of importance to know how the sample corresponds to the target group of which sentiment conclusions are drawn. This problem is known in the field of psychological and sociological research methodology as the external validity of research, which is defined by Shadish et al. ([13], p. 83) as: "… inferences about whether the cause-effect relationship holds over variations in persons, settings, treatments, and outcomes." In this article, the cause-effect relationship is of type product-sentiment or (political) topic-sentiment and the variations in persons, settings, treatments and outcomes between the target group of which the sentiment is measured compared to the available online sample. If the book of Fig. 1 was written for an audience without a background in mathematics, the book should be evaluated by people from such a population, but from a typical sentiment mining report we do not know if this is the case.

Sentiment mining researchers have only recently started to acknowledge the problem of external validity [14]. Wu et al. [15] acknowledge that there is a problem related to customer group representations and propose a visualization of sentiment mining results including customer groups. In response to the many sentiment mining publications based on Twitter data, Mislove et al. [14] found a Twitter population that was highly deviating from the US demographics. Gayo-Avello [16,17] argues that skewness in demographics contributes to failures

* Corresponding author.
E-mail addresses: a.b.j.m.wijnhoven@utwente.nl (F. Wijnhoven), oscarbloemen@gmail.com (O. Bloemen).

0167-9236/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.dss.2013.12.005



★★★★☆ **Excellent book**, June 22, 2012

By **Ben Watson "huntertom"** (huntertom) - See all my reviews

REAL NAME

Amazon Verified Purchase (What's this?)

This review is from: **Data Mining: Practical Machine Learning Tools and Techniques, Third Edition (The Morgan Kaufmann Series in Data Management Systems) (Paperback)**

Great book that would be useful to people with a background in mathematics and programming looking to really take the leap into machine learning. This book has really helped me grasp a lot of the ideas behind the techniques used in ML. A very good book to have for reference and a good read.

Help other customers find the most helpful reviews
Was this review helpful to you?
Yes
No
Report abuse | Permalink
Comment

Fig. 1. Review over Witten et al. (2011) from Amazon.com, accessed August 16, 2012.

of predicting electoral outcomes from social media content and encourages research toward automatic profiling of social media content authors.

This article attempts to fill this gap on the external validity of sentiment mining by taking into account the context of opinion-expressing authors. We elaborate on context extraction methods, following a product-oriented design theory approach [18], which involves first the detection of its kernel theory, next the identification of its meta-requirements, third the listing of meta-designs for solution artifacts, and finally testing the validity of design propositions. The creation of solution artifacts "… relies on existing kernel theories that are applied, tested, modified, and extended through the experience, creativity, intuition, and problem solving capabilities of the researcher" ([19], p. 76). A product-oriented kernel theory provides ideas for meta-requirements and meta-designs that help to solve classes of problems and create classes of artifacts. Next, design propositions describe effective relations between requirements and designs that can be subject to tests [18,20]. If the design propositions can be corroborated, i.e. sufficient evidence is found that the resulting design does what it is required to do, the bias identifier is expected to be reliable and useful for empirically identifying the size of biases in a sample of sentiment-expressions. If the proposed design propositions cannot be corroborated, no statement about the existence of biases in the sample is possible. Section 2 gives a kernel theory of external validity in social sciences from which three possible sentiment-mining biases are derived. Section 3 gives results of a structured literature review to identify meta-requirements and meta-designs for a sentiment mining biases identifier. Section 4 gives tests of the design and empirical propositions. Finally, Section 5 gives the conclusions and implications of this study.

2. Kernel theory: external validity of sentiment mining reports

Shadish et al. ([13], p. 86–90) give five threats to external validity. The first threat (T1) reflects the properties of the sample units, for example the gender and educational level of the people in the sample, and how they relate to the causal relationship. The second threat (T2) relates to differences in treatments. A found relationship might not hold in combination with other treatments or variations of the treatment. Example is a possible payment for participation in an experiment. The third threat (T3) indicates that findings of a specific study cannot be extrapolated to different outcomes. Shadish et al. [13] here give the example of establishing the effectiveness of a medical treatment, which could be measured in quality of life, 5-year metastasis-free survival, or overall survival. These outcomes may differ and thus cannot be easily generalized to each other. The fourth threat (T4) indicates that observations may be biased by specific settings that do not represent the situations over which one wants to generalize. For example, the test of medical drugs may have different results in developed and developing countries due to different health hazards in both. The fifth threat (T5) is related to the way that causal patterns are identified. The paths that explain causal relationships can be different across various settings. For example the financial crisis in The Netherlands may be reinforced by a too high consumption of mortgages, whereas in Greece it is reinforced by poor government budget control.

Application of these five threats to sentiment mining reveals possible problems with sentiment mining results. From threat T1, the first form of possible bias B1 is due to a mismatch in demographic properties of the sample and target audience, e.g. if the researcher is interested in the public opinion of a specific population, the authors of these opinion expressions must reflect this population. The next problem lies in the motivation of posting a review online. Reviews can be written to purposely influence public sentiment, i.e. manipulating the perceived sentiment (B2). An example for such a motivation could be to increase sales for a specific item by posting positive reviews.

Threat T2 introduces a problem related to personal experiences of the author, i.e. the sentiment is biased by specific events (B3). Examples include: a review author with a negative sentiment due to certain problems with an old product that would not occur in the new version, or review authors may develop a generally negative attitude due to conditions without any relevance for the product, like the negative evaluations of a movie after several power outages during its presentation.

Threat T3 relates to the type of information that is extracted with the sentiment mining tool. If the interest is an overall sentiment regarding a product, this should be extracted, but if conclusions are drawn about specific features of the product, generalization toward general sentiments may be invalid. This is a kind of analysis error caused by the non-comparability of aspects or features that are mined. Careful selection of the aspects and features in the mining method therefore is a fundamental task for avoiding external validity problems [9–11].

Threat T4 describes the importance of the research setting when generalizing the findings. In sentiment mining research, the setting of the website(s) from which the reviews are mined could be troublesome for generalization. For instance, mining an online forum for Apple product users to determine sentiments regarding Samsung products is expected to give different results than doing the same on an Android forum. Such platform biases (B4) involve a combination of previously mentioned demographic, manipulation and event biases.

Threat T5 [21] relates to the causal path that links analysis of sentiments to the sentiments of the author (B5), which lies within the applied sentiment mining algorithm. These paths are typically described by features found using a machine learning algorithm. The majority of publications in sentiment mining research concerns with refinements of these algorithms [12].

Table 1 gives an overview of the relations between threats to external validity in social sciences and possible problems in sentiment mining research. For this article, we focus on biases due to demographics, events, and manipulation.

Using the following SCOPUS query ["opinion mining" OR "sentiment analysis" *OR* (*Mining AND* ("social media" *OR* "user generated content" OR reviews OR blog OR forum*)] we found a large set of relevant literature on sentiment mining. The set was made more specific by extending the query with ["external validity" OR generali* OR sample OR noise OR