Decision Support Systems 113 (2018) 86–98

Contents lists available at ScienceDirect

Decision Support Systems

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

By the numbers: The magic of numerical intelligence in text analytic systems[☆]

Richard Gruss[a,*], Alan S. Abrahams[b], Weiguo Fan[c], G. Alan Wang[b]

[a] Department of Management, College of Business and Economics, Radford University, Radford, VA 24142, USA
[b] Department of Business Information Technology, Pamplin College of Business, Virginia Tech, Blacksburg, VA 24061, USA
[c] Department of Management Sciences, Tippie College of Business, University of Iowa, Iowa City, IA 52242, USA

ARTICLE INFO

Keywords:
Text analytics
Information retrieval
Numerical attributes
Defect discovery

ABSTRACT

There is a growing recognition among MIS researchers and practitioners that social media provide a valuable source of business intelligence. Unearthing relevant and useful information among the voluminous postings remains a challenge, however. Automated methods based on text mining have made significant progress in recent years by discovering a variety of new methods and features. This study adds to this stream by introducing a novel text mining procedure centered around numerical expressions contained in text documents. In this method, numerical expressions are extracted, categorized, and binned, and their presence and magnitude are stored as document features. We demonstrate, using a case study from the automotive industry, that numerical expressions can be reliably identified, and that these numerical features enable improvements in document classification. As an extension to this case study, we contribute a decision support system for managing product quality using both textual and numerical attributes.

1. Introduction

The recent surge of academic interest in the use of social media reflects the growing role of customer to customer (C2C) communications as a resource for data-driven business decision making. Researchers are discovering efficient methods for businesses to find, among the vast quantities of social media postings, opportunities for innovation [39], manufacturing defects [2,4,24,39,49], and unanticipated consumer safety issues [46].

Although several approaches for locating high quality content have been explored using social network analysis (SNA), most of the information in social media is expressed as unstructured text, and therefore many promising techniques are contributed by text mining (TM) and natural language processing (NLP). Some text-based methods that have proven effective include term prevalence metrics [2,24,46], latent Dirichlet allocation (LDA) [17,40], and ensemble methods [26].

One token type that is largely ignored in text analytic research is the numeric type. A number by itself conveys very little information, because the lexical token alone does not provide context for understanding its referent or the significance of its magnitude. For example in the two snippets “I was traveling at 25mph …” and “I was going at 15 miles per hour …”, a typical token-based approach would extract only unigrams like “25”, “15”, “mph”, and trigrams like “miles per hour”. These disjoint tokens hold little information value, as “mph” is not associated with “miles per hour” and the relative magnitude of 25 vs. 15 vs. other speed values is not recognized. In contrast, a numeric information extraction approach would recognize that both snippets are distinct expressions of measures of “vehicle speed”, and that the authors were describing travel at “low” speeds (e.g. relative to say 80 mph travel).

This study aims to fill this gap by proposing a procedure for creating a set of numerical features that communicate information about the semantics of numbers. We demonstrate the value of this approach with a case study in automotive defect identification.

When key terms are extracted using prevalence metrics as in [4], numbers sometimes appear in the list of n-grams that are significantly associated with the target class. These numbers are usually removed from curated term lists on the grounds that they communicate little information in isolation. Likewise, important numerical terms might not make it onto prevalent term lists because that particular value appears too infrequently when regarded solely as a lexical token. Furthermore, because the token alone is recognized, and not similar values, new examples would be incorrectly classified if the token value in the new observation is even slightly different from past observations.

[☆] This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

[*] Corresponding author.
E-mail address: rgruss@radford.edu (R. Gruss).

https://doi.org/10.1016/j.dss.2018.07.004
Received 2 March 2018; Received in revised form 30 June 2018; Accepted 31 July 2018
Available online 03 August 2018
0167-9236/ © 2018 Elsevier B.V. All rights reserved.

Consider, “15” and “25” in the example above: these would typically have insufficient prevalence to make it onto the prevalent terms list, as the distinct values are peculiar to each observation. However, once recognized as examples of “low speed”, the semantic concept of “low speed” may indeed have unusually high prevalence in a target class of documents that contain, for example, automobiles with manufacturing defects. The document feature set would include, instead of simply a “most prevalent terms” list, a “most prevalent semantic concepts” list. Furthermore, a new observation where the user is traveling at 10 mph or 20 mph would be correctly recognized as falling into the same category of low speed travel (compared to “10” and “20” being distinct and hitherto unseen tokens, in a conventional token-recognition approach, which would fail to recognize the significance of these values).

This study is an attempt to enrich the document feature set by assigning categories to numbers based on their function in the text. Significant n-grams (strings of words of length n), or “Smoke words,” introduced in [4], are marker terms that are substantially more prevalent in the target class of text documents than in the non-target class. Their original use case was in differentiating online postings indicative of manufacturing defects in automobiles. While our case study specifically involves the automotive industry, we maintain that this procedure is generalizable to any domain with a large variety of meaningful numerical attributes.

Tools that make use of numerical expressions in NLP tasks typically follow a procedure in which entities are first located using Named Entity Recognition (NER), and then numerical attributes are attached based on textual proximity. The goal in this methodology is to build a database of entities with a structured representation. For example, in the sentence, “We tested an all-wheel-drive XLE model as well, which also delivered more than its 24-mpg promise.” [1], “XLE” would be identified as a car model, and the algorithm would need to add “mpg = 24” to that entity. An example of this approach can be found in [7]. Our approach differs in that we first find the numbers, learn their magnitude and units, and create indicators on the level of the social media posting. For example, instead of storing the fact that the XLE has an mpg of 24, we store the fact that a high mpg was mentioned in the posting. Our representation of the posting retains the original extracted value, but in addition, we discretize the value also, as that is informative for defect classification and for slice-and-dice drill-down of the textual dataset by numeric bands (e.g., rapidly finding all postings mentioning high fuel efficiency vehicles).

This paper addresses three main research questions. *First*, can a reasonably-sized set of domain-specific numerical attributes be identified for a given industry? *Second*, can these numbers be processed and interpreted automatically? *Third*, are these numbers useful for general information mining tasks? We demonstrate via a case study that the answer to all of these questions is yes.

Our primary contribution is methodological. We propose a procedure for identifying and classifying a set of domain-relevant numerical attributes with a high level of precision and recall. Furthermore, we demonstrate how these numerical attributes can be combined with key term extraction to achieve improved performance in defect isolation tasks.

The rest of this paper is organized as follows. In Section 2, we provide our theoretical motivation and make the case for generalizability. In Section 3, we review related work in processing numerical attributes. In Section 4, we detail the procedure for numerical attribute extraction and classification. In Section 5, we demonstrate how the procedure works using a case study in the automotive industry, evaluating the effectiveness of the numerical attributes in locating defects in social media postings relative to earlier approaches that were all agnostic to numerical attributes. In Section 6, we further demonstrate the procedure's utility within this case study by proposing a Post Market Defect Surveillance System that offers a dynamic interface for exploring a social media data set using numerical attributes as filters and facets. In Section 7, we discuss limitations and future work. Finally, in Section 8, we discuss our conclusions and implications for research and practice.

2. Theoretical rationale

We contend that by advancing from a strictly lexical treatment of a numerical token to an interpretation of the number's function and magnitude, we are endowing a text analytic system with some degree of semantic understanding. Many attempts to add semantic comprehension to text classifiers have been inspired by findings from cognitive science about how humans process meaning. The methods of artificial intelligence, in many cases, were derived by defining the processes of human intelligence (however narrowly) and approximating them computationally. Examples from natural language processing (NLP) include word sense disambiguation, topic analysis, named entity recognition, and recognizing textual entailment. We argue that our procedure of extracting and binning of numbers would add numerical intelligence to a variety of tasks in several domains, and we make our case from three perspectives: numeracy, specificity, and semantic richness.

Numeracy, which normally develops in childhood [12], is a basic understanding of numbers and their magnitudes. A variety of mental competencies are associated with numeracy, including estimating, ranking, understanding probabilities and making comparisons. People who are numerate are less likely fall prey to the biases that lead to poor decisions [34]. A lack of numeracy has been associated with making poor health decisions [42] and defaulting on a mortgage [16]. An estimation of magnitudes is a fundamental part of human intelligence, and our procedure is an attempt to add this competency to NLP systems.

Another reason why numerical intelligence can aid a variety of NLP tasks is that numbers represent an enhancement of specificity. Human language can be abstract and ambiguous, and so utterances that are specific and concrete provide an opportunity for natural language parsers. When people make reference to numbers, they are producing evidence about a specific case. For example, compare “my car doesn't start on cold mornings” to my “2002 model 56x doesn't start when it falls below 32°.” In addition to establishing their own credibility and competence, the speakers are acknowledging the particularity of their case, and this particularity should be leveraged for its information content.

Our third theoretical basis for generalizability has to do with semantic richness. In cognitive science, “semantic richness” refers to the amount of information associated with a concept [22] and is a function of the variability of that concept's usage and contexts. Three measures of semantic richness are commonly used: 1) number of semantic neighbors (NSN), 2) number of features (NOF), and 3) contextual dispersion (CD). NSN refers to the number of words that are used in a similar context to the focal word, NOF refers to how many different attributes of the concept are available in memory, and CD refers to the number of different contexts in which the word is commonly used.

Several experiments have confirmed that words that are semantically rich are understood more quickly and accurately than those that are semantically impoverished. People are able to perform lexical decision and categorization tasks faster and more accurately for more semantically rich words [35–37,48]. When a word with few semantic neighbors, few features, and few contexts is presented to a person, more effort is required to arrive at an understanding of the word's meaning.

The aspect of semantic richness most applicable to machine learning is the number of features (NOF). The “features” of a concept are simply attributes that are associated with it, such as, for “grapefruit”, < is a fruit > and < is healthy > . [31] found that people are able to list more features of some words than for others. High-NOF words are comprehended more quickly and accurately than low-NOF words [18,48]. An advance in the semantic information supplied to a machine, therefore, is to add features to concepts. In our case study, we take numerical tokens from automobile postings and add features to them. For example, “200” is the lexical token, but we add features “ < is a