Computers in Human Behavior 63 (2016) 132–141

Contents lists available at ScienceDirect

Computers in Human Behavior

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

Full length article

Social media research: The application of supervised machine learning in organizational communication research.

Ward van Zoonen*, Toni, G.L.A. van der Meer

The Amsterdam School of Communication Research ASCoR, University of Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Received 25 November 2015
Received in revised form
10 May 2016
Accepted 12 May 2016

Keywords:
Twitter
Supervised machine learning
Communication research
Content analysis

ABSTRACT

Despite the online availability of data, analysis of this information in academic research is arduous. This article explores the application of supervised machine learning (SML) to overcome challenges associated with online data analysis. In SML classifiers are used to categorize and code binary data. Based on a case study of Dutch employees' work-related tweets, this paper compares the coding performance of three classifiers, Linear Support Vector Machine, Naïve Bayes, and logistic regression. The performance of these classifiers is assessed by examining accuracy, precision, recall, the area under the precision-recall curve, and Krippendorf's Alpha. These indices are obtained by comparing the coding decisions of the classifier to manual coding decisions. The findings indicate that the Linear Support Vector Machine and Naïve Bayes classifiers outperform the logistic regression classifier. This study also compared the performance of these classifiers based on stratified random samples and random samples of training data. The findings indicate that in smaller training sets stratified random training samples perform better than random training samples, in large training sets ($n = 4000$) random samples yield better results. Finally, the Linear Support Vector Machine classifier was trained with 4000 tweets and subsequently used to categorize 578,581 tweets obtained from 430 employees.

1. Introduction

Social media use in organizations is evolving at an unprecedented rate (Treem & Leonardi, 2012). Social technologies may enable employees to communicate more effectively, widen the scope of their work and boost their performance (e.g., Ollier-Malaterre, Rothbard, & Berg, 2013). Likewise, social media can play a crucial role in organizations' post-crises communication (Schultz, Utz, & Goritz, 2011; van Zoonen & van der Meer, 2015), as well as in their efforts to influence evaluations of corporate reputation (Helm, 2011) and to engage in stakeholder dialogues (Lovejoy, Waters, & Saxton, 2012). Some scholars argue that social media are reshaping the nature of the workplace and of work itself (Bucher, Fieseler, & Suphan, 2013), whereas others suggest social media are the new hybrid element in the promotion mix (Mangold & Faulds, 2009). Hence, it is of no surprise that social media is top of the agenda for many practitioners and scholars (Kaplan & Haenlein, 2010).

Organizational communication scholars frequently rely on content analysis research for a wide range of empirical questions. Specifically so, the advent of social media propelled the use of content analysis in organizational communication research as it provides a wealth of well-documented information (e.g., Gallaugher & Ransbotham, 2010; Ki & Nekmat, 2014; McCorkindale, 2010; Rybalko & Seltzer, 2010; van Zoonen, Verhoeven, & Vliegenthart, 2016). The use of content analysis is particularly appealing to researchers because examination of narrative texts - such as social media interactions and blogs - allow the unobtrusive study of organization-public interactions that are otherwise difficult to obtain (Duriau, Reger, & Pfarrer, 2007).

The growth of social media use in the workplace has provided an abundance of data that reflects new media activities and artifacts relevant to organizational communication research (Lewis, Zamith, & Hermida, 2013). In spite of the alluring promise of online data availability, analysis of this information remains arduous. Online data abundance often proves to be 'fool's gold' as: a) researchers often struggle to 'trap' these large data streams, b) large-scale content analyses tend to be unfeasible due to the time-

* Corresponding author. University of Amsterdam, The Amsterdam School of Communication Research *ASCoR*, Nieuwe achtergracht 166, 1018 VW Amsterdam, The Netherlands.
E-mail address: w.vanzoonen@uva.nl (W. van Zoonen).

http://dx.doi.org/10.1016/j.chb.2016.05.028
0747-5632/

consuming and costly realities of human coding methods, and c) computer-assisted content analysis methods are insufficiently institutionalized in organizational communication research (Lewis et al., 2013). In order to alleviate these concerns this study aims to explore the application of an advanced computer-assisted content analysis approach – i.e., supervised machine learning (SML: Grimmer & Stewart, 2013 Hillard, Purpura, & Wilkerson, 2008; Rusell & Norvig, 2002) – that could benefit a variety of research streams in organizational communication.

This study applies a method that is based on SML requiring that a computer learns how to automatically predict content-analytical variables in the corpus of data from a set of human-coded training documents. Notably, in the past SML procedures have predominantly been applied to larger texts such as newspaper articles or speech transcripts (Burscher, Odijk, Vliegenthart, de Rijke, & de Vreese, 2014; Hillard et al., 2008; Scharkow, 2013). Significantly, this study explores the performance of SML to *short* social media messages - i.e., tweets, which are typically limited to 140 characters. Introducing the use of SML for social media content minimizes both the efforts and investments required for content analysis of big data, and helps address substantial issues currently faced by organizational communication scholars and practitioners. The central aim of this study is to answer the research question: To what extent can SML be applied to the content analysis of social media messages?

This study explores the application of SML and aims to advance organization studies by providing a method with which to overcome the limitations associated with both human coding procedures and computer coding procedures. If SML can be used to code social media content a vast array of online information relevant to organizational research awaits empirical analysis.

Moreover, this study explores what type of classifier yields the most reliable coding decisions. The performance of three different classifiers is examined: Linear Support Vector Machine, Naïve Bayes, and standard logistic regression. The algorithms are used to identify the extent to which tweets are related to employees' work. The performance of these classifiers in terms of categorizing social media content is assessed by evaluating key performance indices – i.e., accuracy (AC), recall (RC), precision (PC), the area under the precision-recall curve (AUC), and Krippendorf's Alpha (KA). Additionally the SML procedure is applied to a dataset containing tweets of employees from various organizations. Thereby, this study demonstrates the use of SML in the content analysis of employees' social media messages. In the next paragraph a discussion of the application of content analysis in organizational communication research is provided, followed by an analysis of the case study used to test SML in the coding of social media content. The case study builds on the work by van Zoonen et al., 2016.

2. Content analysis in organizational communication

For organizational communication scholars, central to the value of content analysis is the assumption that content analysis of text and speech provides a replicable methodology to access deep individual or collective structures, such as values, intentions, attitudes, and cognitions (Huff, 1990). That is, ensuring content analysis is applicable to a broad range of organizational phenomena. Content analysis in the field of organizational communication research has included crisis communication (An & Gower, 2009), corporate social responsibility (Campopiano & De Massis, 2015; Cho & Hong, 2009), organization-public communication of nonprofit organizations (Lovejoy & Saxton, 2012; Waters & Jamal, 2011; Waters, 2007; Waters, Burnett, Lamm, & Lucas, 2009), strategic management (Short & Palmer, 2008), and employees' work-related use of personal social media accounts (van Zoonen et al., 2016).

The current era of *Big Data* is both as enticing as it is vexing for communication scholars, as it offers a vast array of information on human and organizational communication, yet the technique and time needed to extract and analyze this data can prove unreasonably longwinded. Nonetheless, the quantity of information available has attracted a wide range of content analysis work in the field of organizational communication. In this thriving branch of research, studies predominantly focus on one specific issue (e.g., Humphreys, Gill, Krishnamurthy, & Newbury, 2013; Small, 2011), event (e.g., Reinhardt, Ebner, Beham, & Costa, 2009) organization or context (e.g., Chew & Eysenbach, 2010). The rationale behind this is a pragmatic one - to reduce the data size and data depth, to make either human coding or dictionary approaches more feasible.

In content analysis research, scholars use a technique to systematically, objectively and quantitatively describe manifest communication content (Berelson, 1952; Lewis et al., 2013). Several approaches to content analysis of social media data have been adopted in communication research. When investigating social media content, scholars often rely on human coding with indicator questions or dictionary-based computer-aided coding.

The most widely used approach in content analysis is human coding, which uses indicator questions often formulated in codebooks. It is often applauded for its systematic rigor and sensitivity towards the subtleties in human language. However, whilst being a reliable method, it is also a highly resource-intensive process. Furthermore, in times where researchers must no longer choose between data size and data depth (since data is abundantly available online) human coding procedures present fundamental challenges to content analysis (Lewis et al., 2013). Researchers are increasingly confronted with 'too much data', forcing them to resort to one of three following tactics: 1) collecting smaller samples by using less sources, reducing the timeframe, or narrowing the context of the study, 2) using random or stratified sampling methods to reduce data size (Riffe, Lacy, & Fico, 2005) or 3) allocating more financial resources (if available) to increase the number of coders hired to carry out the work (Holsti, 1969). Computational methods could offer a solution to some of the sampling and coding limitations of human coding procedures (Lewis et al., 2013).

There are two types of deductive automated coding procedures, dictionary based coding and SML. Deductive automated coding relies on a priori defined categories, in contrast to inductive automated content analysis where categories are automatically derived from the data. One of the most widely used computational coding procedures employed in communication science is dictionary-based computer-aided coding. In such cases, character strings and rules for their combination are defined a priori to code text units into content categories (Krippendorff, 2004). For example, in sentiment analysis of political texts, words such as 'respect,' 'vindication,' and 'cheerfulness' were identified as indicative for positive sentiment, whereas 'insolence,' 'malevolence,' and 'painfulness' indicated negative sentiment (Young & Soroka, 2012). Although this permits the analysis of very large datasets, several drawbacks of dictionary-based approaches cannot be overlooked. For instance, the applicability of a dictionary-approach is limited, as the phenomenon under study needs to be singular; a precondition that is often violated in organizational communication research, as research questions often span across organizational boundaries and require group comparisons and between-subject designs. Consider the following example - in a representative sample of the workforce the use of employees' personal Twitter accounts for work is analyzed. However, in order to code employees' tweets one cannot rely on a dictionary approach, as employees work in a variety of industries and jobs. Thus, creating an a priori set of words and combinations to determine work-related use becomes