Lukoianova, T., & Rubin, V. (2014). Veracity Roadmap: Is Big Data Objective, Truthful and Credible?.
Advances In Classification Research Online, 24(1). doi:10.7152/acro.v24i1.14671

Veracity Roadmap: Is Big Data Objective, Truthful and Credible?

Tatiana Lukoianova
Fisher College of Business
Ohio State University
250 Fisher Hall
2100 Neil Ave
Columbus, OH, USA 43210
vashchilko.1@fisher.osu.edu

Victoria L. Rubin
Language and Information Technology
Research Lab (LIT.RL)
Faculty of Information and Media Studies
University of Western Ontario
North Campus Building, Room 260,
London, Ontario, Canada N6A 5B7
vrubin@uwo.ca

ABSTRACT

This paper argues that big data can possess different characteristics, which affect its quality. Depending on its origin, data processing technologies, and methodologies used for data collection and scientific discoveries, big data can have biases, ambiguities, and inaccuracies which need to be identified and accounted for to reduce inference errors and improve the accuracy of generated insights. Big data veracity is now being recognized as a necessary property for its utilization, complementing the three previously established quality dimensions (*volume*, *variety*, and *velocity*), But there has been little discussion of the *concept* of veracity thus far. This paper provides a roadmap for theoretical and empirical definitions of veracity along with its practical implications. We explore veracity across three main dimensions: 1) objectivity/subjectivity, 2) truthfulness/deception, 3) credibility/implausibility – and propose to operationalize each of these dimensions with either existing computational tools or potential ones, relevant particularly to textual data analytics. We combine the measures of veracity dimensions into one composite index – *the big data veracity index*. This newly developed veracity index provides a useful way of assessing systematic variations in big data quality across datasets with textual information. The paper contributes to the big data research by categorizing the range of existing tools to measure the suggested dimensions, and to Library and Information Science (LIS) by proposing to account for heterogeneity of diverse big data, and to identify information quality dimensions important for each big data type.

Keywords
Big data, veracity, deception detection, subjectivity, credibility, natural language processing, text analytics.

Advances in Classification Research, 2013, November 2, Montreal, QC, Canada.

INTRODUCTION

"Not everything that counts can be counted, and not everything that can be counted counts."
- Albert Einstein

With the Internet producing data in massive volumes, important questions arise with regard to big data as an object or phenomena in itself, and its main characteristics that can support big data-driven discoveries. Do "numbers speak for themselves … with enough data" (Anderson, 2008)? Does big data provide "insights we have never imagined" after mining "masses of data for new solutions and understanding" (Ayshford, 2012)? Or does big data have intrinsic biases, since "data and data sets are not objective; they are creations of human design" (Crawford, 2013)? Big data emerges as the main source and "the heart of much of the narrative literature, the protean stuff that allows for inference, interpretation, theory building, innovation, and invention" (Cronin, 2013, p. 435).

The trade-off in any big data set is between cost and quality of information. Technological developments in the last century have made information one of the most valuable national and private resources, though the main concern was the access costs to information, data[1] gathering, and its sharing (Adams, 1956; Brien & Helleiner, 1980; Mosco & Wasko, 1988; Read, 1979). Today, as volume continues to increase measuring in petabytes and costs continue to decrease, the quality issues of information have become more important than ever before (Hall, 2013). IBM estimates that poor data quality costs US consumers about $3.1 trillion per year and about 27% of respondents in one survey were unsure of how much of their data was inaccurate (2013). "Since much of the data deluge comes from anonymous and

[1] "The difference between data and information is functional, not structural," and as such "data itself is of no value until it is transformed into a relevant form" (Fricke, 2008). However, this paper raises additional issue: low quality data once transformed produces low quality information. Thus, data has to be examined for its truthfulness, objectivity, and credibility to produce corresponding information – truthful, objective, and credible.

ISSN: 2324-9773

unverified sources, it is necessary to establish and flag the quality of the data before it is included in any ensemble" (Dasgupta, 2013).

However, it is only recently that the importance of information quality (IQ) has been recognized, with calls for characterizing big data not only along the three established dimensions, the so-called three "V"s, volume, variety, and velocity, but also along a fourth "V" dimension: veracity (Schroeck, Shockley, Smart, Romero-Morales, & Tufano, 2012). Until recently, the 3Vs, older intrinsic qualities, have led to a 'soup' of data: "content" has been treated like a kind of soup that "content providers" scoop out of pots and dump wholesale into information systems" (Bates, 2002). Still, despite the discussions of the need to examine the veracity of big data, almost no attempts have been made to investigate its nature as a theoretical phenomenon, its main components and the ways to measure it. This is an important limitation of current big data research and practice, since without identifying big data veracity big data-driven discoveries are questionable. This paper attempts to fill this gap.

Veracity goes hand in hand with inherent uncertainty in big data which is predicted to increase rapidly within next two years (Schroek et al 2012). But "despite uncertainty, the data still contains valuable information" (Schroek et al 2012, p. 5). To extract value from big data, information has to be verified to establish its veracity by managing its uncertainty.

Uncertainty management of mainly numeric non-textual data can be done either by "combining multiple less reliable sources" to create "a more accurate and useful data point" or using "advanced mathematics that embraces it [uncertainty], such as robust optimization techniques and fuzzy logic approaches" (Schroeck et al., 2012, p. 5). Uncertainty management of textual data is more complex, since the textual data in general, and especially, from social media "is highly uncertain in both its expression and content" (Claverie-Berge, 2012, p. 3). However, management of uncertainty in textual data gains importance with "the total number of social media accounts" exceeding "the entire global population" (Claverie-Berge, 2012, p. 3).

This paper delineates a roadmap to veracity for textual big data by suggesting ways of managing uncertainty in content and expression. We propose to manage content uncertainty by quantifying the levels of content objectivity, truthfulness, and credibility (OTC), and to manage expression uncertainty by applying Rubin's (Rubin, 2006, 2007) methodology to evaluate sentence certainty. In particular, we argue that quantification of subjectivity, deception and implausibility (SDI) reduces uncertainty with regard to textual data content by providing knowledge about levels of the SDI. The SDI levels are the basis for information verification, and, as such, OTC are the main dimensions of big data veracity. We propose to calculate a big data veracity index by averaging SDI levels. Content characterized by low levels of SDI indicates acceptable veracity, and, therefore, is appropriate for subsequent analysis. On the contrary, content characterized by high levels of SDI needs more cleaning, or in extreme cases cannot be used at all.

We argue that the proposed uncertainty management method for textual big data content and expression increases quality of information and, thereby, improves subsequent analysis by decreasing bias and errors stemming from big data uncertainty. In particular, we reason that high quality big data is objective, truthful, and credible (OTC), whereas big data of low quality is subjective, deceptive and implausible (SDI). Thus, this paper delineates theoretical dimensions of big data veracity, OTC; suggests their potential operationalization; offers a novel quantitative indicator to measure veracity, the big data veracity index[2]; and categorizes currently existing computational linguistics tools, which can be used to measure veracity dimensions.

Blending multidisciplinary research on deception detection, objectivity and credibility with information quality (IQ) in LIS and Management Information Systems (MIS), the paper contributes to information quality assessment (IQA) by adding one more dimension, veracity, to the intrinsic IQ of big data. In particular, we specify two main types of uncertainty in textual big data, expression and content, the effective management of which helps to establish veracity.

The paper is structured in the following way. First, the paper reviews recent literature on information quality, big data, uncertainty, and OTC. The second part theorizes how management of content and expression uncertainty in textual data can contribute to information verification, and thereby, establish big data veracity. The third part suggests ways to operationalize each of the veracity dimensions and develops the big data veracity index. The fourth part identifies and categories each of the existing or potential tools to quantitatively assess veracity dimensions and the overall veracity. The final part sums up our contribution and concludes with practical implications for research and practitioner communities in LIS, classification indexing, text-processing and big data analytics.

LITERATURE REVIEW

Research on IQ defines and assesses information quality based on the usefulness of information or its "fitness for use" by delineating various dimensions along which IQ

[2] Some analytics have called for some sort of "veracity score" measure to assess levels of veracity in big data (Walker, 2013), however, no research has been implemented on it.

ISSN: 2324-9773

can be measured quantitatively (Juran (Juran, 1992; Knight & Burn, 2005; Lee, Strong, Kahn, & Wang, 2002; Stvilia, 2007; Stvilia, Al-Faraj, & Yi, 2009; Stvilia, Gasser, Twidale, & Smith, 2007). One of the four major dimensions of IQ is intrinsic IQ, in which various authors assigned such components as accuracy, believability, reputation, objectivity (Richard Y Wang & Strong, 1996), accuracy and factuality (Zmud, 1978), believability, accuracy, credibility, consistency and completeness (Jarke & Vassiliou, 1997), accuracy, precision, reliability, freedom from bias (DeLone & McLean, 1992), accuracy and reliability (Goodhue, 1995), accuracy and consistency (Ballou & Pazer, 1985), correctness and unambiguousness (Wand & Wang, 1996). However, many of these theories and methodologies cannot be directly applied to the evaluation of big data quality due to the nature and context of big data characterized by inherent uncertainty, especially in textual information (Schroeck et al., 2012). Uncertainty can come from multiple sources such as data inconsistency and incompleteness, ambiguities, latency, deception, as well as model approximations. For the purposes of analyzing big textual data quality, however, uncertainty should be broadly categorized into two main categories: expression uncertainty and content uncertainty (Claverie-Berge, 2012).

Traditionally in LIS, uncertainty has been dealt with in the context of information seeking, for instance, as the basic principle of information seeking (Kuhlthau, 1993), a perceived relevance or potential usefulness of information (Attfield & Dowell, 2003), a cognitive gap (Yoon & Nilan, 1999) and (Dervin, 1983). In textual data, expression uncertainty and ambiguity are encoded in verbal expressions, like hedging and qualifying statements (Rubin, 2007, 2010). This interpretation of the concept of expression uncertainty, as analyzed within natural language processing (NLP), has to do with an intentional language ambiguity mechanism: people encode variable assessments of the truth of what is being stated. Uncertainty, in this sense, is "a linguistic and epistemic phenomenon in texts that captures the source's estimation of a hypothetical state of affairs being true" (Rubin, 2010). The work on identification of factuality or factivity in text-mining (e.g., Morante & Sporleder, 2012; Saurí & Pustejovsky, 2009, 2012) stems from the idea that people exhibit various levels of certainty in their speech and that these levels are marked linguistically (e.g., *maybe, perhaps* vs. *probably and for sure*) and can be identified with NLP techniques (Rubin, 2006; Rubin, Kando, & Liddy, 2004; Rubin, Liddy, & Kando, 2006).

For example, (Rubin et al., 2006) empirically analyzed a writer's (un)certainty, or epistemic modality, as a linguistic expression of an estimated likelihood of a proposition being true. An analytical framework for certainty categorization was proposed and used to describe how explicitly marked (un)certainty can be predictably and dependably identified from newspaper article data (Rubin, 2006). The certainty identification framework serves as a foundation for a novel type of text analysis that can enhance question-and-answering, search, and information retrieval capabilities.

Much has been written in LIS on credibility assessment and a variety of ways and checklist schemes to verify the credibility and stated cognitive authority of the information providers (e.g., Fogg & Tseng, 1999; Rieh, 2010). Rieh (2010) summarizes the historical development of the credibility research in such fields as psychology and communication, and provides a recent overview of credibility typologies in LIS (e.g., source credibility, message credibility, and media credibility) and HCI (e.g., computer credibility: presumed credibility, reputed credibility, surface credibility, and experienced credibility). With automation in mind, Rubin and Liddy (2006) defined a framework for assessing blog credibility, consisting of 25 indicators in four main categories: blogger expertise and offline identity disclosure; blogger trustworthiness and value system; information quality; and appeals and triggers of a personal nature. Later, Weerkamp and de Rijke (2008; 2012) estimated several of the proposed Rubin and Liddy's indicators and integrated them into their retrieval approach, ultimately showing that combining credibility indicators significantly improves retrieval effectiveness.

The concept of separating subjective judgments from objective became of great interest to NLP researchers and gave rise to a currently active area in NLP – sentiment analysis and/or opinion mining – which is concerned with analyzing written texts for people's attitudes, sentiments, and evaluations with NLP and text-mining techniques. Rubin (2006) traces the roots of subjectivity/objectivity identification work in NLP to Wiebe, Bruce, Bell, Martin, and Wilson (2001) who developed one of the first annotation schemes to classify and identify subjective and objective statements in texts. Prior to this work on subjectivity, Rubin (2006) continues, an NLP system needed to determine the structure of a text – normally at least enough to answer "Who did what to whom?" (Manning & Schütze, 1999). Since early 2000s the revised question was no longer just "Who did what to whom?" but also "Who thinks what about somebody doing what?" For a comprehensive overview of the field of opinion-mining/sentiment analysis, see Pang and Lee (2008) and Liu (2012)).

Another prominent body of research literature of interest to big data quality assessment is that of deception detection. Emerging technologies to identify the truthfulness of written messages demonstrates wide-range problems related to deceptive messages and importance of deception detection in textual information. Deception is

ISSN: 2324-9773

prominently featured in several domains (e.g., politics, business, personal relations, science, journalism, per (Rubin, 2010) with the corresponding user groups (such as news readers, consumers of products, health consumers, voters, or employers) influenced by decreased information quality. However, the IQ research seems to undervalue the role of deception in improving IQ (Knight & Burn, 2005; Lee et al., 2002; Stvilia et al., 2007). Several successful studies on deception detection have demonstrated the effectiveness of linguistic cue identification, as the language of truth-tellers is known to differ from that of deceivers (e.g., Bachenko, Fitzpatrick, & Schonwetter, 2008; Larcker & Zakolyukina, 2012).

We discuss uncertainty, subjectivity, credibility, and deception in conjunction and in the context of big data IQ assessment, to establish big data veracity.

THEORY

We argue that big data can possess different features and characteristics, which affect its quality. As any object, the features of big data can vary across many dimensions. Therefore, depending on its origin, data processing technologies, and methodologies used for data collection and scientific discoveries, big data can have more/less biases, and various other information quality (IQ) features, which need more/less human-computer interactions for scientific discoveries to produce viable solutions. Big data has no value unless it can be effectively utilized and proper utilization of big data depends on recognizing and accounting for those IQ features, which help to reduce inference errors and improve the accuracy of generated insights. These features include inherent content and expression uncertainty, which can undermine big data veracity.

The goal of this paper is to extend the IQA methodology and framework by theoretically conceptualizing and operationalizing big data veracity. The theory builds on research in MIS, LIS, and computational linguistics by explicitly describing expression and content uncertainty along with their components as they contribute to veracity and overall IQ We propose to use three components of content uncertainty – subjectivity, deception and implausibility (SDI), along which we can verify information for its veracity.

Due to inherent uncertainty in big data, veracity has become one of the critical factors in creating value from the standard three "V" dimensions: volume, variety, and velocity (Schroeck et al., 2012). IBM defines veracity as the fourth dimension of big data, which specifically deals with data in doubt, and refers to "the level of reliability associated with certain types of data" including "truthfulness, accuracy or precision, correctness" (IBM, 2013; Schroeck et al., 2012). IBM suggest some direct ways of tackling veracity by "creating context around the data", for example, "through data fusion, where combining multiple less reliable sources creates a more accurate and useful data point, such as social comments appended to geospatial location information" (Schroeck et al., 2012, p. 5). However, IBM and many others lack more generalizable ways of characterizing and assessing big data veracity. This paper attempts to fill this gap.

Each of the traditional big data dimensions, volume, velocity and variety (Figure 1), could be measured quantitatively with a reasonable degree of accuracy.

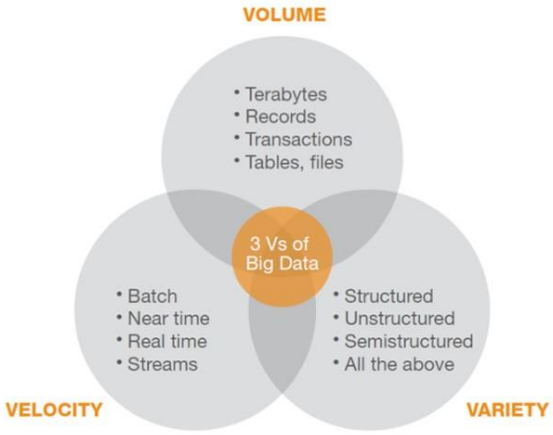

Figure 1. Three Standard Intrinsic Dimensions of Big Data (Claverie-Berge, 2012).

The fourth dimension, veracity, however, is a more complex theoretical construct with no agreed upon ways of measuring it, especially, for non-numeric textual big datasets (Figure 2).

We argue that by decoding uncertainty from verbal expressions and content in textual data, such uncertainty can be identified and diminished, which can improve big data veracity. This is because uncertainty generates not only ambiguities, but also potential factual inconsistencies (Auer and Roy 2008). So, to define and measure veracity, we need to delineate the main sources of expression and content uncertainty, SDI, producing variations in veracity levels. We argue that SDI increase uncertainty of textual big data, and as such lead to the decline in veracity.

Thus, we propose to define three main theoretical veracity dimensions: objectivity, truthfulness, and credibility, CTO. Each of these dimensions characterize various big data problems (as in Schroeck et al. (2012)), and thereby can decrease big data quality along with its value. For example, deception detection is a way of identifying whether verbal expressions are truthful or not as well as whether overall content is truthful or not.

ISSN: 2324-9773

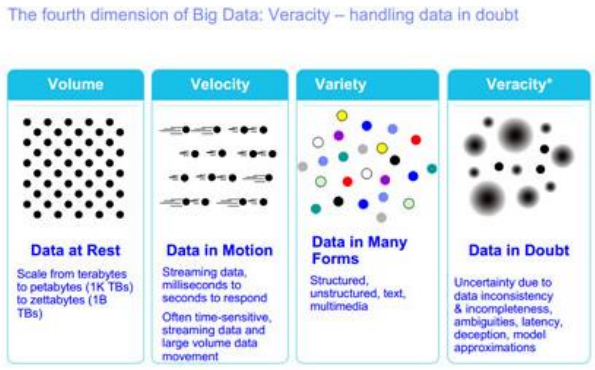

Figure 2. Four dimensions of big data now include Veracity (Claverie-Berge, 2012).

We suggest defining veracity across three dimensions: 1) objectivity, 2) truthfulness, and 3) credibility. Figure 3 visualizes the conceptual space of three primary orthogonal dimensions, objectivity, truthfulness, credibility, since they capture different aspects of textual information. The dimensions intersect in the center and the nebula represents a certain degree of variability within the phenomena that together constitute the big data veracity. [Secondary dimensions (of lesser concern in textual data, and thus, in this paper) are presented in dotted lines]. All three dimensions reduce "noise" and potential errors in subsequent inferences from the textual big data due to minimization of bias, intentional misinformation, and implausibility.

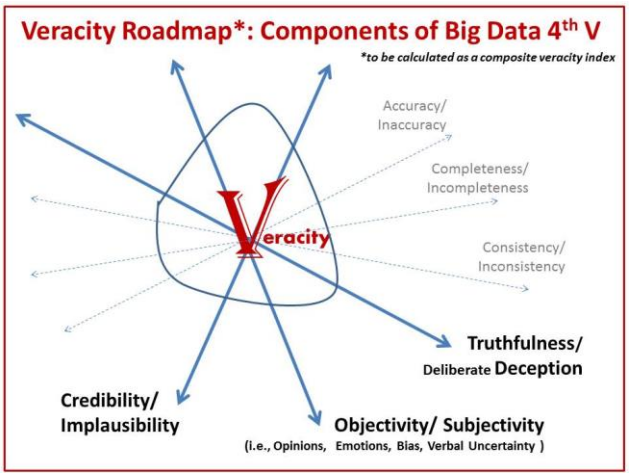

Figure 3. Conceptualization of the Components of Big Data Veracity

Explicit or implicit objectivity relies on information sources (McQuail, 2010), or refers to understanding of information (Hjørland, 2007). For example, many news agencies and various official sources of information might have explicit biases, whereas, the objectivity of personal blogs/social media being is less obvious, and thus, most likely more subjective.

Deception refers to intentional misinformation, or a deliberate attempt to create a false belief or a false conclusion (e.g., Buller & Burgoon, 1996; Zhou, Burgoon, Nunamaker, & Twitchell, 2004). The implausibility[3] of textual information refers to data quality, capability, or power to elicit disbelief; it undermines data validity and weakens trust in its content, rendering data potentially unusable and any related scientific discoveries – questionable (Roget's 21st Century Thesaurus, 2013).

To delineate various dimensions of veracity, our paper draws on the main concerns with regard to quality of information in the disciplines that either produce large amounts of textual information (media) or manage and curate digital information (LIS, MIS). Since these disciplines have developed a detailed understanding of the main issues with data and information quality, we can utilize this knowledge to define both expression and content uncertainty along with the veracity dimensions. We also rely on media theory with regard to the objectivity/subjectivity and credibility/implausibility dimensions, since media (both social and traditional, e.g., blogs and digital news online) is one of the three[4] main sources of big data. The credibility/ implausibility and truthfulness/ deception veracity dimensions are grounded in NLP and in LIS's primary concern with information authority control.

We first discuss each type of uncertainty managing (expression, content) which helps to establish veracity.

Expression uncertainty

Expression uncertainty – not to be confused with the concept of overall doubt in data, or content uncertainty – refers to linguistic marking of the strength of the content-generators' convictions. "Facts and opinions can be expressed with a varying level of certainty in news writing as well as other genres such as scientific literature and belle-lettre. Some writers consciously strive to produce a particular effect of certainty, either due to training or explicit instructions, and others may do it inadvertently. Many statements have evident traces of such writers' behavior. Some writers' levels of certainty seem constant throughout the text and can be unnoticed by the reader. Those of others' fluctuate from statement to statement and shifts between attributed sources and the writer's opinions." *(Rubin, 2006, p. 5)*

We argue that textual information is filled with linguistic markers that could help to manage not only expression

[3] The term 'implausibility' is used here synonymously with 'improbability' and 'unreasonableness', and as an antonym to 'credibility' (per Roget's 21st Century Thesaurus (2013)) suitable in the context of big data use, interpretation, and comprehension.

[4] "Big data is often boiled down to a few varieties including social data, machine data, and transactional data." (Smith, 2013).

ISSN: 2324-9773

uncertainty, but also content uncertainty.

Content Uncertainty

We argue that content uncertainty management can be improved by categorization into three main components based on the main sources of content ambiguities: subjectivity, deception and implausibility. Thus, independently of the context, textual information can vary across three veracity dimensions, objectivity, truthfulness, and credibility. This contrasts to Mai (2013), who argued that "information quality is context-dependent, and can only be assessed and understood from within specific situations and circumstances" (p. 675).

Objectivity/Subjectivity Dimension

The subjectivity/objectivity of meaning can arise in textual information from the writer, the reader, or neither of them, *objet trouvé* (Hirst, 2007, p. 3). Objectivity, "especially as applied to news information," is "the most central concept in media theory with relation to information," since "objectivity is a particular form of media practice … and also a particular attitude to the task of information collection, processing and dissemination" (McQuail, 2010, p. 200). "Objectivity has to deal with values as well as with facts and the facts also have evaluative implications" (McQuail, 2010, p. 201). This is an "information producer" view of information objectivity.

Within philosophical discussions, "objectivity" – in one of the prominent uses of the term – is typically associated with ideas such as truth, reality, reliability, and the nature of support a particular knowledge-claim: "*Objective knowledge* can designate a knowledge-claim having, roughly, the status of being fully supported or proven. Correspondingly, "subjective knowledge" might designate some unsupported or weakly supported knowledge-claim" *(Internet Encyclopedia of Philosophy: A Peer-Reviewed Academic Resource, 2013).*

From the "information consumer" point of view, the objectivity/subjectivity dimension relates to how information is understood (Hjørland (2007)):

"1. The objective understanding (Observer independent, situation independent). Versions of this view have been put forward by, for example, Parker, Dretske, Stonier, and Bates. Bates' version implies: Any difference is information.

2. The subjective/situational understanding. Versions have been put forward by, for example, Bateson, Yovits, Spang-Hanssen, Brier, Buckland, Goguen, Hjørland. This position implies: Information is a difference that makes a difference (for somebody or for something or from a point of view). What is information for one person in one situation needs not be information for another person or in another situation. This view of information as a noun is related to becoming informed (informing as a verb). Something is information if it is informative—or rather, something is information when it is informative" (Hjørland, 2007, p. 1449).

Objectivity though, is different from truth, since objectivity is only one version of truth with truth being a broader notion than objectivity (McQuail, 2010). Therefore, in the definition of veracity, we differentiate between objectivity/ subjectivity and truthfulness/ deception dimensions.

Deception/Truthfulness Dimension

Deception in written communication represents an information quality (IQ) problem by intentionally and knowingly creating a false belief or false conclusion on the part of the sender in the mind of the receiver of the information (e.g., Buller & Burgoon, 1996; Zhou et al., 2004). Passing the deception detection test can verify the source's intention to create a truthful impression in the readers' mind, supporting the trustworthiness and credibility of sources. On the other hand, failing the test immediately alerts the user to potential alternative motives and intentions and necessitates further fact verification.

For big data, deception can grow along with the amount of data itself, thereby increasing its uncertainty. "With the massive growth of text-based communication, the potential for people to deceive through computer-mediated communication has also grown and such deception can have disastrous results," (Fuller et al. 2011, p. 8392). Identification of deception in big data helps to diminish content uncertainty, and, therefore, deception should constitute one of the main dimensions of the veracity.

Credibility/Implausibility Dimension

Media theory also differentiates objectivity from credibility, both of which have become intrinsic parts of journalism with credibility in this context having the same meaning as believability in 1950s (Johnson & Wiedenbeck, 2009). "Credibility is, after all, the most important thing a communicator has. A communicator in the news media who lacks credibility probably has no audience" (Severin & Tankard, 1992, p. 28). Tseng and Fogg (1999) elaborated that, in a more sophisticated view, credibility is defined as a perceived quality of trustworthiness and expertise, simultaneously evaluated. Trustworthiness refers to goodness or morality of the source and can be described with terms such as well-intentioned, truthful, or unbiased. Expertise refers to perceived knowledge of the source and can be described with terms such as knowledgeable, reputable, and competent (Tseng & Fogg, 1999). Expertise is also of prime concern in authority evaluations work such as by Conrad, Leidner, and Schilder (2008). "The most credible

ISSN: 2324-9773

information is found in those perceived to have high levels of trustworthiness and expertise though "[t]rustworthiness and expertise are not always perceived together" (Rieh, 2010, p. 1338).

The concept of trust is often used in everyday language, and communication in making trustworthiness decisions. Hardin (2001) noticed a pervasive conceptual slippage that involves a misleading inference from the everyday use of trust: many ordinary-language statements about trust seem to conceive trust, at least partly, as a matter of behavior, rather than an expectation or a reliance. In relation to big data and Web information, trust is an assured reliance on the character, ability, strength, or truth of trusted content ("Merriam-Webster Online Dictionary," 2009).

Two credibility components, trustworthiness and expertise, are essential to making credibility (i.e., believability) judgments about trustworthiness (i.e., dependability) of entities or information, regardless of whether such judgments are expressed lexically with a vocabulary of trust as being trustworthy (i.e., dependable), or credible (i.e., believable).

METHODOLOGY: OPERATIONALIZATION OF VERACITY DIMENSIONS AND THE BIG DATA VERACITY INDEX

The paper proposes to operationalize each veracity dimension by describing how OTC are measured with either existing computational tools or potential ones, since the dimensions are mutually exclusive and reflect different aspects of big data veracity. The paper contributes to the big data research by categorizing the range of existing tools to measure the suggested dimensions. Objectivity-subjectivity variation in many ways depends on its context, since context determines the types of linguistic cues used to express objective or subjective opinions (Hirst, 2007). To quantify deception levels in big data, we propose to use the existing automated tools on deception detection (see overview in (Rubin & Conroy, 2012; Rubin & Lukoianova, Forthcoming; Rubin & Vashchilko, 2012). For credibility assessment, we propose to use blogs that contain trust evaluation of published content or entire websites.

For the purposes of operationalizing veracity and its dimensions, it is useful to focus not on the concept of information per se, but rather on the meaning that information carries, as in computational linguistics. Even though Mai (2013) argues that "information quality is context-dependent, and can only be assessed and understood from within specific situations and circumstances" (p. 675), it seems that, for big data, information context is important only for the choice of the most appropriate tools to reduce uncertainty and establish veracity.

Tools for Detecting Subjectivity, Opinions, Biases

Many of the recent computational linguistics tools automate and assist in interpretations such as, "automatic classification of the sentiment or opinion expressed in a text; automatic essay scoring" (Hirst, 2007, p. 7). The development of such tools has been gaining popularity in recent years reflecting the attention to subjective information and ways to distil its interpretation. This also indicates the existence of subjective information, which needs to be differentiated across variations in subjectivity.

Sensitivity to nuance thus requires, for any particular utterance in its context, knowing what the possible alternatives were. Clearly, this kind of analysis requires both complex knowledge of the language and complex knowledge of the world. The latter may be arbitrarily hard — ultimately, it could imply, for example, a computational representation of a deep understanding of human motivations and behavior that even many people do not achieve (Hirst, 2007, p. 8). (See Rubin (2006) for a description of the development of subjectivity software)..

The resulting tools can, for instance, identify political biases, pool opinions on a particular product from product-reviews, or create more effective cross-document summaries for automatic news aggregators[5]. Subjective content, however, does not necessarily discount the validity of the information, since subjective statements (those from a particular angle) can still be informative, truthful, and valid.

Deception Detection Tools

Automated deception detection is a cutting-edge technology that is emerging from the fields of NLP, computational linguistics, and machine learning, building on years of research in interpersonal psychology and communication studies on deception.

The main two reasons for using automation in deception detection are to increase objectivity by decreasing potential human bias in detecting deception (reliability of deception detection), and improve the speed in detecting deception (time processing of large amounts of text), which is especially valuable in law enforcement due to time-sensitivity (Hutch et al 2012). However, Hutch et al (2012) demonstrate that computational tools might provide conflicting findings on the direction of the effect of the same linguistic categories on the level of deception in textual (non-numeric) information.

The majority of the text-based analysis software uses

[5] The challenge for NLP-enabled tools remains in scaling up to the big data volume and managing the constantly incoming stream (its velocity). These tools often require time consuming deep-parsing, data enrichments, and multiple passes through the data prior to making automated classification decisions (e.g., whether a product was liked or not, based on its reviews, and if not, why).

ISSN: 2324-9773

different types of linguistic cues. Some of the common linguistic cues are the same across all deception software types, whereas other linguistic cues are derived specifically for the specialized topics help to generate additional linguistic cues. For example, Moffit and Giboney's (2012) software calculates the statistics of various linguistic features present in the written textual information (number of words, etc.) independently on its content, and subsequently these statistics can be used for classification of the text as deceptive or truthful. The language use represented by linguistic items changes under the influence of situational factors: genre, register, speech community, text or discourse type (Crystal, 1969).

The automation of deception detection in written communication is mostly based on the linguistic cues derived from the word classes from the Linguistic Inquiry and Word Count (LWIC) (Pennebaker, Francis, & Booth, 2001). The main idea of LWIC coding is text classification according to truth conditions. LWIC has been extensively employed to study deception detection (Hancock, Curry, Goorha, & Woodworth, 2007; Mihalcea & Strapparava, 2009; Vrij, Mann, Kristen, & Fisher, 2007).

Vrij et al. (2007) compared the LWIC approach to manual coding to detect deception, and concluded that the manual analysis is better than the LWIC-used computational analysis. However, the most recent analysis of automated deception detection with software to detect fake online reviews demonstrated a significant improvement of computational approaches over human abilities to detect deception (Ott, Choi, Cardie, & Hancock, 2011). The goal of Ott et al. (2011) was to identify fake reviews of products and services on the Internet. Several software programs (Chandramouli and Subbalakshmi 2012, Ott et al. 2011, Moffit and Giboney 2012) were evaluated in our previous work (Rubin and Vashchilko 2012). The majority of the software offers on-line evaluation tools without algorithm provision (Chandramouli and Subbalakshmi 2012), or with the provision of API (Ott et al. 2011, Moffit and Giboney 2012), and customizable dictionaries (Moffit and Giboney 2012). For discussion of advantages and disadvantages of various approaches and the comparative evaluation details of the software capabilities (Rubin & Vashchilko, 2012). Further analysis of similar deception detection tools is needed to determine which of them are particularly suitable for detection deception in big data to establish its veracity.

Credibility Tracking Tools

The opinion-mining approach of analyzing combined personal experiences, evaluations, and recommendations, in essence, provides an alternative source of information for a reputation-based knowledge structure for a trust-system, and as such can serve as a basis to measure the credibility/implausibility dimension of veracity. If an entity (person, organization) or information is trusted by multiple opinion-holders, it can be inferred to be trustworthy, even though the individual entities are not necessarily trusted. The power of multiple low-trust entities providing similar judgments independently should not be undermined. For instance, Gil and colleagues (2006) suggest that if a high-trust entity contradicts the judgments of multiple independent low-trust entities, the credibility of the information provided by such a high-trust entity may be questioned.

The success of the system largely depends on its ability to identify and retrieve a subset of relevant blogs. The difficulty in obtaining such relevant blogs with a simple query (e.g., "*trust OR credibility*") is what motivated current work, as a step toward constructing sufficiently informative queries to selecting an appropriate subset of data to be further analyzed. Particularly, by looking at the inventory of words that frequently and consistently collocate with the terms in questions and their definitional and derivational extensions, we can identify differences and similarities in general language use, predict what roles the surrounding terms can play in retrieved blog opinions, and refine the queries accordingly.

Mutual information (MI)-based collocation analysis[6] of nouns and verbs most frequently occurring with *trust* and *credibility* identified distinct lexico-semantic spaces as used in the Corpus of Contemporary American English (COCA) COCA is a large freely available online corpus representing contemporary use of the language, 1990-2008. At the time of data collection and analysis, the corpus contained 387 million words of text, about 20 million words a year (Davies, 2009). The three concepts of interest to us as a seed for a reputation system – *trust*, *credibility*, and *trustworthiness* – collocate with *integrity*. *Honesty* collocates with both *trust* and *trustworthiness*; *confidence* – with *trust* and *credibility*; and *competence* and *character* – with *trustworthiness* and *credibility*. This implies that *credibility* collocations are, perhaps, of most use for discovering the abstract notions of reasons and justifications for credibility judgments, e.g., *competence, accuracy, and prestige. Trust* has its own set of justifications, e.g., *respect, goodwill, decency;* and possible opinion-holders or targets, *e.g., leadership, government, parents.*

Overall, this corpus linguistics approach – as a shallow parsing method (that is limited to part-of-speech

[6] Mutual Information (MI) is a method of obtaining word association strength. The MI between two words, $word_1$ and $word_2$ is defined as:

$$MI(word_1, word_2) = \log_2 \left(\frac{p(word_1 \& word_2)}{p(word_1)\ p(word_2)} \right)$$

In this formula, p(*word₁ & word₂*) is the probability that *word₁* and *word₂* co-occur. "If the words are statistically independent, the probability that they co-occur is given by the product p*(word₁)* p*(word₂)*. The ratio between p*(word₁ & word₂)* and p*(word₁)* p*(word₂)* is a measure of the degree of statistical dependence between the words." (Turney & Littman, 2003).

.

ISSN: 2324-9773

knowledge about each word in the corpus) achieves its goal of revealing significant relationships around the central terms, which is conceptually insightful, as well as practically applicable to retrieving a rough pool of relevant texts in unseen data. The limitations of this approach are that it is still "a bag-of-words" method that ignores syntactic structures (e.g., in terms of phrase, clause, and sentence boundaries); it ignores the roles each word perform semantically (e.g., an argument or a recipient of an action); it ignores negation (simple use of particle "not"). However, the above-mentioned collocations were identified as potential seed terms suitable for a social media credibility-monitoring system.

Veracity Index

The paper offers to combine the three measures of veracity dimensions into one composite index, the veracity index. The tree main components of veracity index, OTC, are normalized to the (0,1) interval with 1 indicating maximum objectivity, truthfulness and credibility, and 0, otherwise. Then, the big data veracity index is calculated as an average of OTC, assuming that each of the dimensions equally contributes to veracity establishment. However, the authors acknowledge that each dimension can contribute to the overall quality of big data to a different degree, and can be assigned different weights in the big data veracity index. This can happen, if one of the veracity dimensions, say deception in insurance claims, can be of outmost importance for the subsequent analysis, and, inherently, all insurance claims are subjective, so subjectivity dimension might not needed at all to establish data veracity.

Thus, this newly developed veracity index provides a useful way of assessing systematic variations in big data quality across datasets with textual information. Different combinations of these three dimensions, e.g., being objective, truthful, and credible could be seen in multiple examples and are not rare. Therefore, the paper suggests capturing not only the variation across these three dimensions separately, but also overall quality variation evaluated by a composite index[7].

DISCUSSION

In the last few years, conceptual tools dealing with language accuracy, objectivity, factuality and fact-verification have increased in importance in various subject areas due to rising amount of digital information and the number of its users. Journalism, online marketing, proofreading and political science, to name a few. For example, in political science *Politifact* (albeit based on man-powered fact-checking) and *TruthGoggles* sort true facts in politics helping citizens to develop better understanding of politicians statements (Rubin and Conroy, 2012). McManus's (2009) *BS Detector* and Sagan's (1996) *Baloney Detection Kit* help readers to detect fraudulent and fallacious arguments, as well as check the facts in news of various kinds, economic, political, scientific. In proofreading, *Stylewriter* and *AftertheDeadline* help users to identify stylistic and linguistic problems related to their writings. These tools use not only linguistic cues to resolve expression uncertainty problems, but also experts' opinions, and additional necessary sources to establish the factuality of events and statements, which helps to resolve content uncertainty. For an overview of related automation and annotation efforts, see (Morante & Sporleder, 2012; Sauri & Pustejovsky, 2009; Sauri & Pustejovsky, 2012).

Considering several known deception types (such as falsification, concealment and equivocation, per Burgoon and Buller 1994), we emphasize that the deception detection tools are primarily suitable for falsification only. For a recent review and unification of five taxonomies into a single feature-based classification of information manipulation varieties, see Victoria L. Rubin and Chen (2012). Certain types of deception strategies cannot be spotted automatically based on underlying linguistic differences between truth-tellers and liars. For instance, concealment is a deceptive strategy that requires careful fact verification, likely to be performed by humans regardless of the state-of the-art in automated deception detection.

Recently developed software that resolve expression and content uncertainty by detecting deception, subjectivity, and perhaps implausibility in textual information are potential future venues for research in big data information quality assessment. Several deception detection tools we have identified can be considered ready-to-use IQA instruments for assessment of each veracity dimensions as well as overall big data veracity index. Since truthfulness/deception differs contextually from accuracy and other well-studied components of intrinsic information quality, the inclusion of truthfulness/ deception in the set of IQ dimensions has its own contribution to the assessment and improvement of IQ.

Little is known about the applicability of various automated deception detection tools for written communication in various subject areas. The tools became available to public in the last two years with the predominant methodology of text classification into deceptive or truthful based on linguistic cue statistics.

Three concepts – *trust*, *credibility*, and *trustworthiness* – collocate with *integrity*, an additional construct rarely emphasized in academic literature. *Honesty* collocationally overlaps with *trust* and *trustworthiness*; *confidence* unites *trust* and *credibility; competence* and

[7] The index could be helpful to identify those parts of big dataset that are of lower quality for their subsequent exclusion, if the quality of the entire dataset can be significantly improved.

ISSN: 2324-9773

character interlock with *trustworthiness* and *credibility*.

From the systems point of view, the retrieved data is intended as an input to an opinion-mining prototype that analyzes, extracts, and classifies credibility judgments and trust evaluations in terms of their opinion-holders, targets, and justifications in specific areas, such as health care, financial consulting, and real estate transactions. Thus, the described above collocation analysis helps the appropriate construction of query to retrieve those blogs that contain trust evaluations and credibility assessment. As such, the retrieved blogs will provide necessary information regarding the complement evaluation of the credibility of some sources and identification of their objectivity.

The objectivity/subjectivity of the opinions is accessible with the recently developed computational tools for sentiment analysis, opinion mining and opinion identification, overviewed above. See recent overviews in Pang and Lee (2008) and Liu (2012).

Practical Implications

Data-mining textbooks typically advise that about 80-90% of the human effort should be allocated to the process of manual data preparation, tabulation, and specifically data cleaning. We see a similar process needed for big data analysis and pattern discovery to support decision making. The era of big data calls for automated (or semi-automated) approaches to data evaluation, cleaning, and quality assurance. The three intrinsic qualities of the data – its volume, velocity and variety – preclude purely manual data analysis, yet human involvement is important in setting the parameters for computational tools and analytics. The age of big data seems to be driving the rise of big data analytics and many wonder where it leaves library professionals that were trained to deal with individual information bearing objects one at a time, giving each their full attention and time to quality assessment and often extensive commentary.

As of fall 2013, big data analysts are in high demand, being actively sought after, hired and trained. In this rush to re-qualify and reach for new skills, the questions we need to ask is what LIS and adjacent fields (e.g., NLP) have to offer in this newly titled profession given the big data size, mobility, variety and inherent 'noise' and quality uncertainty. We argue that library and information professionals (classifiers, cataloguers, indexers, database managers, and other types of technical services in LIS) are best positioned to transition to these roles of big data analysts to support and complement the big data analytics processes by a) transferring the traditional LIS understandings of managing large data sets such as those collected in libraries catalogues and databases; and if needed, b) acquiring additional expertise in text analytics, text-mining and automated classification. It may be no longer feasible to read, analyze, index, classify, or fact-check every single information bearing object individually (such as a list of purchase transactions or blog observations), but what still applies in this context is the attention to the 'big picture' (e.g., trends and patterns), the attention to detail (e.g., noticing suspicious instances in batches), classification principles (e.g., creating exhaustive and mutual exclusive classes by which to sort data, automatically or not). With proper training, information professionals should be able to manage computational tools, provide meaningful support and develop further methodologies for sorting high and low quality data as part of data preparation, evaluation and information quality assessment in huge constantly evolving datasets.

CONCLUSIONS

Ninety percent of all big data was created in the last two years (Yu, 2012). "For big data, 2013 is the year of experimentation and early deployment" with organizations still struggling "to figure out ways to extract value from big data, compared to last year when governance-related issues were the main challenge" (Yu, 2013). Big data can have value only when its veracity can be established, and, thereby, the information quality confirmed. "Developing a generalizable definition for dimensions of data quality is desirable. … Where one places the boundary of the concept of data quality will determine which characteristics are applicable. The derivation and estimation of quality parameter values and overall data quality from underlying indicator values remains an area for further investigation" (Richard Y. Wang, Kon, & Madnick, 1993). Textual big data veracity depends on effective management of inherent content and expression uncertainty, which manifests itself in subjectivity, deception and implausibility (SDI). By assessing the levels of SDI, textual big data veracity can be evaluated along each of its proposed dimensions, truthfulness, objectivity, and credibility, or in general, by calculating big data veracity index. This paper categorizes existing tools for assessing each of the veracity dimensions to resolve content uncertainty and suggest using Rubin's (2006, 2007) methodology to resolve expression uncertainty.

REFERENCES

Adams, Scott. (1956). Information - a national resource. *American Documentation (pre-1986), 7*(2), 71.

Attfield, Simon, & Dowell, John. (2003). Information seeking and use by newspaper journalists. *Journal of documentation, 59*(2), 187-204.

Ayshford, Emily (2012). The Data Age. *McCormick Magazine*. http://www.mccormick.northwestern. edu/magazine/fall-2012/data-age.html

Bachenko, Joan, Fitzpatrick, Eileen, & Schonwetter, Michael. (2008). *Verification and implementation of language-based deception indicators in civil and criminal narratives.* Paper presented at the Proceedings of the

ISSN: 2324-9773

22nd International Conference on Computational Linguistics-Volume 1.

Ballou, Donald P, & Pazer, Harold L. (1985). Modeling data and process quality in multi-input, multi-output information systems. *Management science, 31*(2), 150-162.

Bates, Marcia (2002). After the dot-bomb: getting web information retrieval right this time. *First Monday, 7*(7).

Brien, Rita Cruise, & Helleiner, G. (1980). The political Economy of Information in a Changing International Economic Order. *Internat. Organization, 34*(4), 445-70.

Buller, David, & Burgoon, Judee (1996). Interpersonal Deception Theory. *Communication Theory, 6*(3),203-42.

Burgoon, Judee, & Buller, David. (1994). Interpersonal Deception: III. Effects of Deceit on Perceived Communication and Nonverbal Behavioral Dynamics. *Journal of Nonverbal Behavior, 18*(2), 155-184.

Claverie-Berge, Isabelle (2012). Solutions Big Data IBM. http://www05.ibm.com/fr/events/netezzaDM_2012/ Solutions_Big_Data.pdf

Conrad, Jack G., Leidner, Jochen, & Schilder, Frank. (2008). *Professional Credibility: Authority on the Web*. Paper presented at the 2nd Workshop on Information Credibility on the Web (WICOW08), Napa, CA.

Crawford, Kate (2013). The Hidden Biases in Big Data. *Harvard Business Review*. http://blogs.hbr.org/ cs/2013/04/the_hidden_biases_in_big_data.html

Cronin, Blaise (2013). Editorial. *Journal of the American Society for Inform. Science & Technology, 63*(3), 435–6.

Crystal, David. (1969). *What is linguistics?* Edward Arnold.

Dasgupta, Arup (2013). Big data: The future is in analytics. *Geospatial World, April*.

Davies, Mark. (2009). The Corpus of Contemporary American English (COCA): 387 million words, 1990-present. Retrieved 10 February 2009, from http://www.americancorpus.org.

DeLone, William H, & McLean, Ephraim R. (1992). Information systems success: the quest for the dependent variable. *Information systems research, 3*(1), 60-95.

Dervin, Brenda. (1983). An overview of sense-making research: Concepts, methods, and results to date. *Paper presented at the The Annual Meeting of the International Communication Association, Dallas, TX*.

Fogg, B. J., & Tseng, Hsiang. (1999). *The elements of computer credibility.* Paper presented at the SIGCHI conference on Human factors in computing systems: the CHI is the limit, Pittsburgh, Pennsylvania, United States.

Fricke, M. (2008). The knowledge pyramid: a critique of the DIKW hierarchy. *Journal of Inf. Science, 35*(2), 131-42.

Goodhue, Dale L. (1995). Understanding user evaluations of information systems. *Manag. Science, 41*(12), 1827-44.

Hall, Kathleen (2013). Data quality more important than fixating over big data, says Shell VP. *Computerweekely.com*. http://www.computerweekly.com/news/2240186887/Data-quality-more-important-than-fixating-over-big-data-says-Shell-VP

Hancock, Jeffrey T, Curry, Lauren E, Goorha, Saurabh, & Woodworth, Michael. (2007). On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. *Discourse Processes, 45*(1), 1-23.

Hardin, Russel. (2001). Conceptions and Explanations of Trust. In K. S. Cook (Ed.), *Trust in Society* (pp. 3-39). New York, NY: Russell Sage Foundation.

Hirst, Graeme (2007). Views of text-meaning in computational linguistics: Past, present, and future. In G. Dodig-Crnkovic & S. Stuart (Eds.), *Computing, philosophy, and cognitive science: The Nexus and the Liminal*. Newcastleupon-Tyne: Cambridge Scholars Press.

Hjørland, Birger. (2007). Information: Objective or subjective/situational? *Journal of the American Society for Information Science & Technology, 58*(10), 1448-56.

Jarke, M. , & Vassiliou, Y. . (1997). *Data warehouse quality: a review of the DWQ project*. Paper presented at the Conference on Information Quality, Cambridge, MA.

Johnson, K. A., & Wiedenbeck, S. (2009). Enhancing Perceived Credibility of Citizen Journalism Web Sites. *Journalism & Mass Communication Quarterly, 86*(2), 332-348.

Juran, Joseph M. (1992). *Juran on quality by design: the new steps for planning quality into goods and services*: SimonandSchuster. com.

Knight, Shirlee-Ann, & Burn, Janice M. (2005). Developing a framework for assessing information quality on the World Wide Web. *Informing Science: International Journal of an Emerging Transdiscipline, 8*(5), 159-172.

Kuhlthau, Carol. (1993). A principle of uncertainty for information seeking. *Journal of Documentation, 49*(4), 339-55.

Larcker, David F, & Zakolyukina, Anastasia A. (2012). Detecting deceptive discussions in conference calls. *Journal of Accounting Research, 50*(2), 495-540.

Lee, Yang, Strong, Diane, Kahn, Beverly, & Wang, Richard. (2002). AIMQ: a methodology for information quality assessment. *Information & Management, 40*(2), 133-46.

Liu, Bing. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Lang. Tech., 5*(1), 1-167.

Mai, Jens-Erik. (2013). The quality and qualities of information. *Journal of the American Society for Information Science and Technology, 64*(4), 675-688. doi: 10.1002/asi.22783

Manning, Christopher D, & Schütze, Hinrich. (1999). *Foundations of statistical natural language processing* (Vol. 999): MIT Press.

McQuail, D. (2010). *McQuail's Mass Communication Theory*: SAGE Publications.

Merriam-Webster Online Dictionary. (2009). Retrieved 10 February 2009, from www.merriam-webster.com

Mihalcea, Rada, & Strapparava, Carlo. (2009). *The lie detector: Explorations in the automatic recognition of deceptive language.* Paper presented at the Proceedings of the ACL-IJCNLP 2009 Conference Short Papers.

Morante, Roser, & Sporleder, Caroline. (2012). Modality and negation: An introduction to the special issue. *Computational linguistics, 38*(2), 223-260.

Mosco, V., & Wasko, J. (1988). *The Political Economy of Information*: University of Wisconsin Press.

Ott, Myle, Choi, Yejin, Cardie, Claire, & Hancock, Jeffrey T. (2011). Finding deceptive opinion spam by any stretch of the imagination. *arXiv preprint arXiv:1107.4557*.

Pang, Bo, & Lee, Lillian. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval, 2*(1-2), 1-135.

Pennebaker, James W, Francis, Martha E, & Booth, Roger J. (2001). Linguistic inquiry and word count (LIWC): A computerized text analysis program. *Mahwah (NJ), 7*.

ISSN: 2324-9773

Read, William H. (1979). Information as a National Resource. *Journal of Communication, 29*(1), 172-178.

Rieh, Soo Young. (2010). Credibility and Cognitive Authority of Information. In B. M & M. N. Maack (Eds.), *Encyclopedia of Library and Information Sciences, Third Edition* (pp. 1337 - 1344). New York: Taylor and Francis Group.

Roget's 21st Century Thesaurus. (2013). Credibility. (n.d.). http://thesaurus.com/browse/credibility

Rubin, Victoria L. (2006). Identifying certainty in texts. *Unpublished Doctoral Thesis, Syracuse University, Syracuse, NY*.

Rubin, Victoria L. (2007). Stating with Certainty or Stating with Doubt: Intercoder Reliability Results for Manual Annotation of Epistemically Modalized Statements. *Proceedings of NAACL HLT 2007, Companion Volume*, 141-144.

Rubin, Victoria L. (2010). Epistemic modality: From uncertainty to certainty in the context of information seeking as interactions with texts. *Information Processing & Management, 46*(5), 533-540. doi: 10.1016/j.ipm.2010.02.006

Rubin, Victoria L., & Conroy, Neil. (2012). Discerning truth from deception: Human judgments and automation efforts. *First Monday, 17*(3).

Rubin, Victoria L., Kando, Noriko, & Liddy, Elizabeth D. (2004). *Certainty categorization model.* Paper presented at the AAAI spring symposium: Exploring attitude and affect in text: Theories and applications, Stanford, CA.

Rubin, Victoria L., & Liddy, Elizabeth D. (2006). *Assessing Credibility of Weblogs.* Paper presented at the AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs.

Rubin, Victoria L., Liddy, Elizabeth D, & Kando, Noriko. (2006). Certainty identification in texts: Categorization model and manual tagging results *Computing attitude and affect in text: Theory and applications* (pp. 61-76): Springer.

Rubin, Victoria L., & Lukoianova, Tatiana. (Forthcoming). Truth and Deception at the Rhetorical Level. *The Journal of the American Society for Information Science and Technology*.

Rubin, Victoria L., & Vashchilko, Tatiana. (2012). Extending Information Quality Assessment Methodology: A New Veracity/Deception Dimension and its Measures. *The Proceedings of the American Society for the Information Science and Technology, 49*, 1-6.

Rubin, Victoria L. , & Chen, Yimin (2012). Information Manipulation Classification Theory for LIS and NLP. *ASIST 2012, October 28-31, 2012, Baltimore, MD, USA*.

Saurí, Roser, & Pustejovsky, James. (2009). FactBank: a corpus annotated with event factuality. *Language resources and evaluation, 43*(3), 227-268.

Saurí, Roser, & Pustejovsky, James. (2012). Are you sure that this happened? assessing the factuality degree of events in text. *Computational Linguistics, 38*(2), 261-299.

Schroeck, Michael , Shockley, Rebecca , Smart, Janet , Romero-Morales, Dolores, & Tufano, Peter (2012). Analytics: The real-world use of big data. How innovative enterprises extract value from uncertaindata. www.stthomas.edu/gradsoftware/files/BigData_RealWorldUse.pdf

Severin, W.J., & Tankard, J.W. (1992). *Communication theories: origins, methods, and uses in the mass media*: Longman.

Smith, Heather (2013). Big data FAQs – a primer. biblog.arcplan.com/2012/03/big-data-faqs-a-primer/

Stvilia, Besiki. (2007). A model for ontology quality evaluation. *First Monday, 12*(12).

Stvilia, Besiki, Al-Faraj, Abdullah, & Yi, Yong Jeong. (2009). Issues of cross-contextual information quality evaluation—The case of Arabic, English, & Korean Wikipedias. *Libr. & Inf. Science Research, 31*(4), 232-9.

Stvilia, Besiki, Gasser, Les, Twidale, Michael B., & Smith, Linda C. (2007). A framework for information quality assessment. *Journal of the American Society for Information Science and Technology, 58*(12), 1720-33.

Turney, P., & Littman, M.L. (2003). Measuring Praise and Criticism: Inference of Semantic Orientation from Association. *ACM Transactions on Information Systems, 21*(4), 315-346.

Vrij, Aldert, Mann, Samantha, Kristen, Susanne, & Fisher, Ronald P. (2007). Cues to deception and ability to detect lies as a function of police interview styles. *Law and human behavior, 31*(5), 499.

Walker, Michael. (2013). Data Veracity - Data Science Central.pdf. *Data Science Central*. http://www.datasciencecentral.com/profiles/blogs/data-veracity

Wand, Yair, & Wang, Richard Y. (1996). Anchoring data quality dimensions in ontological foundations. *Communications of the ACM, 39*(11), 86-95.

Wang, Richard Y, & Strong, Diane M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems, 12*(4), 5-5.

Wang, Richard Y. , Kon, Henry B. , & Madnick, Stuart E. . (1993). Data Quality Requirements Analysis and Modeling. *The Ninth International Conference of Data Engineering, Vienna, Austria, April 1993, December* (TDQM-92-03).

Weerkamp, Wouter, & de Rijke, Maarten. (2012). Credibility-inspired ranking for blog post retrieval. *Information Retrieval, 15*(3-4), 243-277.

Weerkamp, Wouter, & de Rijke, Maarten (2008). Credibility improves topical blog post retrieval. *Proceedings of ACL08: HLT, page 923-931, Columbus, Ohio. Association for Computational Linguistics, Association for Computational Linguistics*.

Wiebe, Janyce, Bruce, Rebecca, Bell, Matthew, Martin, Melanie, & Wilson, Theresa. (2001). *A corpus study of evaluative and speculative language.* Paper presented at the Proceedings of the Second SIGdial Workshop on Discourse and Dialogue-Volume 16.

Yoon, Kyunghye, & Nilan, Michael S. (1999). Toward a reconceptualization of information seeking research: focus on the exchange of meaning. *Information Processing & Management, 35*(6), 871-890.

Zhou, Lina, Burgoon, Judee K., Nunamaker, Jay F., & Twitchell, Doug. (2004). Automating Linguistics-Based Cues for Detecting Deception in Text-Based Asynchronous Computer-Mediated Communications. *Group Decision and Negotiation, 13*(1), 81-106.

Zmud, Robert W. (1978). An empirical investigation of the dimensionality of the concept of information *. *Decision Sciences, 9*(2), 187-195.

ISSN: 2324-9773