ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal
Regular Issue, Vol. 9 N. 4 (2020), 5-30
eISSN: 2255-2863
DOI: https://doi.org/10.14201/ADCAIJ202094530

Methods for Assessing, Predicting, and Improving Data Veracity: A survey

Fatmah Y. Assiri

University of Jeddah, College of Computer Science and Engineering. Department of Software Engineering, Jeddah, Saudi Arabia
fyassiri@uj.edu.sa

KEYWORD

improvement; prediction; social media; internet of things (IoT); web applications.

ABSTRACT

Data is an essential part of smart cities, and data can play an important role in veracity; estimation; decision making processes. Data generated through web applications and devices utilize the Internet of Things (IoT) and related technologies. Thus, it is also important to be able to create big data, which has historically been defined as having three key dimensions: volume, variety, and velocity. However, recently, veracity has been added as the fourth dimension. Data veracity relates to the quality of the data. Any potential issues with the quality of the data must be corrected because low-quality data leads to poor software construction, and ultimately bad decision making. In this work, we reviewed the existing literature on related technical solutions that address data veracity based on the domain of its application, including social media, web, and IoT applications. The challenges or limitations and related gaps in existing work will be discussed, and future research directions will be proposed to address the critical issues of data veracity in the era of big data.

1. Introduction

The usage of data is a critical or essential component in the formation of smart cities. Further, in a variety of applications the usage of data is used as a basis for decision making, and an essential part of business and organizational success. Data is collected through many applications. For example, web applications collect structured and unstructured data, and the Internet of Things (IoT) devices that are embedded in smart homes collect energy consumption data. Similarly, smart technology in personal mobile devices collect vital signs data and provide self-monitoring systems (Elloumi et al., 2019). The collection of personal data from a wide variety of sources allows for the creation of big data, which is described in terms of three dimensions: volume, variety, and velocity. Additionally, more recent

definitions of big data include veracity as a fourth dimension, as it relates to verifying the quality of big data.

Data veracity does not have a unified definition. The Merriam-Webster dictionary defines it as "conformity with truth or facts." In contrast, some researchers have defined veracity in terms of the level of data uncertainty due to data inconsistency and incompleteness (Debattista et al., 2015). McArdle and Rob defined veracity in terms of authenticity, precision, and reliability of collected data (McArdle and Kitchin, 2016). More simply, others have defined it as data correctness (Agarwal et al., 2016), trustworthiness (Lin et al., 2015), integrity, availability, completeness, and consistency between the data and its sources (Herrera et al., 2019; Patgiri and Ahmed, 2016; Moyne and Iskandar, 2017; Batista et al., 2017). Overall, it can thereby be defined as a process of assigning markers of quality, and then determining the quality and correctness of a data set.

Potential quality issues related to data must be corrected, given that the implementation of low-quality data can lead to poor software development and bad decision-making. Meta Group has reported that 41% of all projects that depend on data-warehousing fail due to poor data quality (Zaparniuk et al., 1995). In 2013, IBM estimated that the annual cost of poor data quality was approximately $3.1 trillion US. According to IBM, the percentage of uncertainty in the data reached 8% by the end of 2015 (Wibowo and Sandikapura, 2019).

Many fields, including but not limited to machine learning, blockchain, and crowdsourcing, have been adapted to solve data veracity issues. This paper aims to review existing work, to identify the challenges described within the existing work, the limitations of that work, and related gaps in the data, and to propose future directions for research to ensure the veracity of data. This research is needed since it has not received the required level of academic attention since entering the era of big data. We classify existing work based on the application domains of existing work, including social media, web, and IoT. Social media uses textual and temporal data in addition to account information to classify tweets and their sources. On the other hand, the web consists of applications that analyze web data to improve data veracity. IoT consists of all applications utilize sensors that are embedded in different devices to collect an individual's data and then assess and predicts data veracity. Existing work aims to assess, predict, or improve data veracity. There are two review papers which have studied data veracity from specific aspects. For example, research carried out by Lozano et al. (Lozano et al., 2020) reviewed research papers that focused on the assessments of the veracity of online data in social media and open sources. Another review paper, conducted by by Eembi (Jamil et al., 2015), reviewed research papers that are related to data veracity as it applies to digital news portals. This study provided statistics of existing research papers related to the aim of the study without providing any technical details. However, to the best of our knowledge, no prior work has reviewed research articles focused solely on technical solutions for different types of data from different domains.

1.1. Purpose and problem Statements

The purpose of this paper is to review existing work to summarize existing technical solutions, which have been researched as it relates to the ability to assess, predict, and improve data veracity. This will be analyzed based on the applications' domain in the reviewed paper, including social media, the web, and IoT. Social media applications used textual and temporal data, the web consists of textual data and its metadata, and IoT consists of sensor data. Thus, we describe the challenges of applying the existing research to practical problems and identifying future research directions in this area. Therefore, the research questions that we are studying are as follows:

- What are the technical solutions that were utilized to assess, predict, and improve data veracity for different applications according to existing literature?
- What are the challenges or short comings of the solutions provided in the existing research, and what are the potential future research directions to improve data veracity?

Thus, the main contributions of this paper are the classification of existing work based on the applications' domain (social media, web, and IoT), the description of the technical solutions that have been utilized to address data veracity, the comparison of the performance of existing methods in terms of the computational and communication costs, and the discussion of challenges, gaps and future directions to improve the veracity of big data.

1.2. Paper outline

The remaining sections of this paper are organized as follows. Section 2 explains data veracity definitions from the perspective of the applications' domains, and Section 3 explains the big data framework and maps existing work related to applications and their corresponding domains. Then, metrics that are used in the evaluation of existing work are defined in Section 4. Section 5 explains the technical solutions that have been proposed for social media applications. Research studies that have been performed to address data veracity for web applications are summarized in Section 6. The methods used to manage data veracity for IoT applications are presented in Section 7. Finally, the challenges of the presented work and future research directions are explained in Section 8.

1.3. Review Methodology

To conduct a systematic review, we followed the guidelines outlined by Kitchenham and Charters (Kitchenham and Charters, 2007) and (García Holgado et al., 2020). We first formulated the research questions, as stated in Section 1.1. We identified criteria to determine the work that would be included in our study. Then, we planned our research to identify search keywords and relevant databases. Papers that directly address the include/exclude aspects were downloaded and reviewed. We first checked each papers' titles, keywords, and abstract. If the information was relevant, then more details were checked by reading the full paper.

We searched different online databases: IEEE Xplore, ACM Digital Library, Science Direct, Scopus, and Web of Science to find the papers in the selected libraries. We conducted a search in each of these platforms using the following keywords: data, veracity, assessment, estimation, prediction, enhancement, improvements, and automation. We created different combinations of the listed keywords to find all the relevant papers. For example, possible rearrangements of the search terms were as follows: data veracity, assessment of data veracity, and data veracity assessment tools. We classified existing studies based on the applications' domains that related to veracity assessment, estimation, prediction, and improvement of different types of data. Papers classification is as shown in Figure 1.

Different libraries were searched for research papers, and then they were classified as relating to social media applications, web applications, and IoT applications. We reviewed approximately thirteen papers from IEEE Xplore, eight papers from ACM, seven from Elsevier, eight from Springer, and nine from other publishers, such as Science Direct. In addition, we used the remaining research papers for veracity definitions, technical details and to explain challenges related to specific areas. The statistics for these papers are shown in Figure 1.

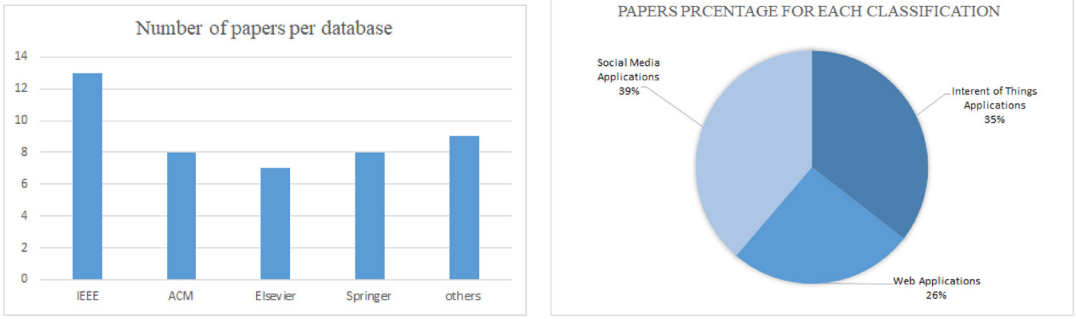

Figure 1: Papers classification.

2. Definition

Data veracity is related to the quality of the data and has been defined as the fourth dimension of big data. There is no specific definition for data veracity in the literature. Ramachandramurthy et al. (Ramachandramurthy et al., 2015) related veracity to information quality. Others have defined data veracity in terms of quality dimensions such as accuracy, confidence, completeness, data volume, and timeliness (Batini et al., 2009), (Klein and Lehner, 2009). In social media applications, data defined veracity in terms of data correctness/genuineness (Agarwal et al., 2016), (Ma et al., 2015), (Wu et al., 2015), (Kwon et al., 2017), (Singh et al., 2019), (Devi et al., 2020). Correspondingly, Oliveria and Giasemidis defined data veracity in terms of trustworthiness (Olivieri et al., 2017), (Giasemidis et al., 2016), and Paryani defined it as text ambiguity (Paryani et al., 2017). However, the focus was the identification of rumors and fake news. On the other hand, in the context of IoT applications, veracity related to both data usability and quality (Liu et al., 2018), and authors focused on data uncertainty (Jeffery et al., 2006), (Diao et al., 2009), (Rodríguez and Servigne, 2013), (Chen and Jiang, 2014) and the reputation of sources in terms of trustworthiness (Lin et al., 2015), (Jagadish et al., 2014).

In the context of web applications, veracity has been defined differently. Lozano et al. referred to data veracity as trust, reliability, and credibility (Lozano et al., 2015). Similarly, Lukoianova and Rubin (Rubin and Lukoianova, 2013) identified data verity in three dimensions: objectivity, truthfulness, and credibility. Moreover, data veracity has been used to ensure authorized access to data (Yin and Kaynak, 2015; Kepner et al., 2014). In addition, different metrics were identified to represent the veracity of web data. Some have related data veracity to the confidence level (i.e., integrity, availability, completeness, consistency, and accuracy) of the data and its sources (Berti-Équille, 2015; Herrera et al., 2019; Patgiri and Ahmed, 2016; Moyne and Iskandar, 2017; Batista et al., 2017). Debattista measured veracity in terms of dereferenceability and consistency (Debattista et al., 2015). In the database area, veracity has been related to data inconsistencies, duplicates, and missing or incomplete data (Berti-Equille and Borge-Holthoefer, 2015).

Figure 2: Big Data Framework.

3. System Model and Applications

The Big data applications framework consists of a data collection layer, a processing layer, a modeling layer, a service layer, and visualisation layers (Tekiner and Keane, 2013), in addition to data and system layers. Figure 2 was adapted from the work of Tekiner and Keane (Tekiner and Keane, 2013). Data collection layers gather data from different sources and store them for application in the following step. Then, the processing (aka. pre-processing) layer integrates collected data, makes any required transformations, such as normalization, and analyzes data to gain insights or output related to patterns in the data. The following layer is a modeling layer, in which the data is mined, and the model is built and evaluated. This allows for the targeted application of data, or use of methods to access results, which then can be visualised by the user in a user-friendly design using specified tools.

The focus of our work is to study the state-of-the-art work that focuses on the validity of collected data, which we argue is part of the pre-processing step. In other words, the data should be validated before it used in model building. Big data applications are widely used in different domains for different purposes. For better review, data veracity techniques and requirements were mapped to applications and their corresponding domains (Table 1). Based on our review, we classified existing works into three main domains: social media, the web, and IoT. In this paper we review, classify, and compare data veracity techniques in order to discuss current challenges and provide future directions to improve data veracity.

Table 1: Big Data Application, its corresponding Domains, and Data Veracity Requirements and Techniques.

Domain	Application	Veracity Requirements	Veracity Techniques
Social Media	Fact checking;	Tweets' text, temporal data, account information, and propagation structure;	Classification algorithms, particle swarm optimization (Kumar et al., 2019), deep learning (Singh et al., 2019), breadth-first search tree structure and spanning tree (Devi et al., 2020), crowdsourcing techniques (Agarwal et al., 2016);
	Text ambiguity;	Tweets' content and topic definition in terms of words;	Topic modeling and natural language processing (Paryani et al., 2017);
Web	Content and expression classification; Linked data;	Expression text and time; Availability of source information;	Natural language processing (Rubin, 2006), (Rubin and Lukoianova, 2013), ensemble methods (Berti-Équille, 2015); Reservoir sampling (Debattista et al., 2015), bloom filter (Debattista et al., 2015);
	Health information classification;	Reliable medical sources and text phrases;	Deep learning (Samuel and Zaiane, 2018);
	Fact checking;	Fake and genuine news text;	Emotion analysis (Tarmizi et al., 2019);
	Access Authorization;	Static and dynamic data;	Cryptography (Kepner et al., 2014);
	Open data;	Data and its metadata;	Crowdsourcing (McArdle and Kitchin, 2016);
	Data exchange systems;	Data providers, data buyers, and data qualifiers collaborations;	Token-based crowdsourcing (Wibowo and Sandikapura, 2019);
	System failure detection;	System components;	Failure mode and effect analysis (Herrera et al., 2019);
Internet of Things	Unmanned aircraft systems;	Uunmanned aerial vehicle (UAV), sensors embedded in vehicles, and historical data;	Compare to UAV data (Li et al., 2020);
	Data collection via sensors;	Parcking historical information and temporal data;	Machine learning algorithms (Ren et al., 2020), influence model (Aman et al., 2014);
	Mobile cloud computation	Collection of individual's data via mobile features;	Category-based context-aware and recommendation incentive reputation mechanism (Lin et al., 2017; ?);
	Rank data sources;	Uncertainty level and score ;	Automated tool (Amini et al., 2016);
	Sensor data uncertainty ;	Raw uncertain data ;	Random sample filter (Jeffery et al., 2006), probabilistic model (Diao et al., 2009), parallel algorithms (Chen and Jiang, 2014; ?), and aggregation scheme (Sanyal and Zhang, 2018).

4. Evaluation Metrics for Data Veracity Techniques

In this work, we reviewed existing techniques that were proposed to assess, predict, and improve data veracity.Some studies focused on the data collection layer, while others focused on modeling layers. One of the biggest challenges or big data is data volume which also impacts the time required for data cleaning, processing, classifying, and validation. In addition, many of big data applications time-consuming and computationally expensive (Kaisler et al., 2013). Thus, to compare between the proposed approaches for assessing, predicting, and improving data veracity, we defined two metrics: computational cost and communication cost that are adapted from Zhou et al. (Zhou et al., 2016) and Oguz et al. (Oguz et al., 2015).

Computational costs are the amount of time an algorithm/process takes to run. As a result, it is a function of input size. As the size of inputs increase, the time will increase. This also increases the computational costs in response.

Communication costs are the amount of communication that is needed between parties to solve a problem; it is computed in terms of bytes transferred. It is expressed as a function of data volume. As such, when data volume increases, the communication cost also increases.

Note that input size, which we also referred to as data volume, is measured in terms of data set size as it is implemented in the experiment of the considered work. Tables 2, 3, and 4 consists of dataset size and data format information; what data will be provided depends on its availability in the study (some did not provide information about the size of the dataset used in the study).

5. Social Media Applications

Social media has a large impact on transferring news, especially fake news related to critical issues. However, credible sources are limited, which makes them unavailable to all users as fake news. There are many works that have tackled this problem, which are called *fact checking*. Predictions have been widely used to predict the veracity of social media data since it is one of the main available sources of big data. Data veracity has been studied in terms of data genuineness (rumor vs. non-rumor). Features, such as tweet's text, user, and propagation behavior, have been extracted to classify rumor and non-rumor tweets (Ma et al., 2015; Wu et al., 2015), and others have used the tweets' paths to classify them (Kwon et al., 2017).

Wu et al. proposed a hybrid support vector machine (SVM) classifier that accounts for a tweet's propagation structure in addition to its semantic features (Wu et al., 2015). A propagation tree is constructed to differentiate between tweets. The propagation tree consists of nodes in which each node represents a user as a normal user or an opinion leader (aka. influencer) along with the message text, number of followers, friends, and reposts for the opinion leader. In the study, a modified random walk graph kernel was used to measure the similarities between trees. In total, 23 features were extracted, and a normal radial basis function was used to measure the distance between them. Some features, such as the topic type, search engine, and user type, were proposed in this work. The classifier was evaluated on Sina Weibo data, which is a Chinese microblogging platform equivalent to Twitter. First, to evaluate the effectiveness of using a graph kernel function, the authors removed the function during the model training. The results showed that using a graph kernel function improved the accuracy. In addition, the proposed SVM was compared to other state-of-the art classifiers. As a result, the authors

found that the hybrid SVM improved the accuracy of the rumor detection within 24 hours following the rumor initiation.

Ma et al. (Ma et al., 2015) proposed a dynamic series-time structure (DSTS) model that predicts rumors based on text features that are collected over a period of time. Thus, it differentiates between rumor and nonrumor text, and in response, it gives a better prediction. First, intervals and time stamps were generated for each event, where the interval was the amount of time, in hours, for the intervals' length, and the timestamp was the index timestamp for a specific text. A feature vector was created for the time stamps that capture the feature for each time interval and the changes between consecutive intervals. In the study, three feature categories were identified: content based, user-based, and propagation-based. For each post, the latent Dirichlet allocation (LDA) model was used to compute the topic-distribution, and a sentiment lexicon and emotion lexicon were used to compute the average sentiment score. The experiment results using Twitter and Sina Weibo data sets showed improvements in the early detection of rumors over time using the presented model.

The veracity of Twitter data has also been predicted in terms of the trustworthiness score (Giasemidis et al., 2016). Rumors features are classified into user-based, content-based, and network-based (Ma et al., 2015; Wu et al., 2015), (Giasemidis et al., 2016). However, the utilization of a user's past behaviors was first introduced in the referenced work. In addition, it took a long time to analyze prior work to determine if the veracity was predicted (an average of three hours and twenty minutes), while the presented work achieved a higher accuracy in an average time of one hour and fifty minutes. The authors' novelty also relied on time-series features to predict the veracity sooner and on building a visualization tool that is user-friendly. Data was collected from Twitter and labeled. Then, the linguistic inquiry and word count (LIWC) were used to extract the linguistic features. The extracted features were user related, message related, and network related. In addition to the time series feature, rumors were divided into 20 time series; in each interval, features were extracted for all tweets starting from the rumor. For the feature selection, the forward selection deterministic wrapper method was used. Only three features were statistically significant: the fraction of tweets for the rumor, users' occupation, and users' followers. The results found that the random forest algorithm obtained the highest precision, and it provided higher accuracy (77%) than previous studies in a shorter time period.

Kwon et al. investigated rumor characteristics over different time windows and determined the best feature selection for early detection (Kwon et al., 2017). Tweets were extracted for different periods of time starting from three days after tweet initialization to 7, 14, 28, and 56 days of the tweets' mobilization. Four sets of features were selected: user-related, linguistic characteristics, spreading network, and temporal features. Data related to the rumors were collected from snopes.com and urbanlegends.about.com, and genuine news was collected from news media. To compare the effect of each feature, random forest algorithm along with three-fold cross-validation was used. The results determined that users' features were effective for rumor classification in the first three days; however, linguistic features had the same effectiveness through different time periods. Network features had poor performance in the early period, whereas temporal features had good performance for a long period. The authors proposed two algorithms for rumor classification. The first algorithm used all of the features; it had low performance during early time periods, but the performance improved after 28 days. Then, for early rumor classification, the authors proposed an algorithm that used only the user and linguistic features. This model had poorer performance compared to the use of all features. However, the performance decreased as the time period increased. This finding is due to the ignorance of the temporal and network features, which improved the classification process over time.

Olivieri et al. proposed the need to assess data veracity and used as a model the resource description framework (RDF) vector (Olivieri et al., 2017). A specified ontology is connected to the meaning of a word; each meaning is classified to help query additional evidence. Then, a vector is generated using classification algorithms (logistic regression, random forest, support vector machine, and neural networks) to predict the trustworthiness of the online input data. The result of the data on movie reviewers showed that using additional information about the domain of interest along with a knowledge base improved the data veracity in comparison with state-of-the-art factchecking methods.

Paryani et al. proposed an entropy-based model to measure the veracity of the Twitter data in terms of text ambiguity (Paryani et al., 2017). The proposed model depends on the bag-of-words distribution of the content and definition. Topic modeling is the model that is used to analyze large unbalanced data to determine the topics; the latent Dirichlet allocation (LDA) is used to extract topics from text based on word occurrences in documents. Then, the proposed model used Shannon's mathematical theory of communication (for more information, see (Shannon, 2001)), which measures text ambiguity.

Shannon's entropy depends on keyword probabilities for those words that define the topic. A lower the resulting entropy value, the higher the indicated level of veracity. The proposed model was evaluated on three datasets related to flu, food poisoning, and politics. They compared the proposed model to two other veracity models: the objectivity, truthfulness, and credibility (OTC) model and the diffusion, geographic and spam indices (DGS) model. The entropy measure can be used as a distance value between the text and the topics or as a value to describe which topics are best defined by its keywords. The results found that all of the models were effective in measuring the veracity, with slight differences; however, the effectiveness of the model depended on the domain of the topic.

Rumor Gauge is a system that was designed to automatically predict the veracity of real-time rumors from Twitter (Vosoughi et al., 2017). The proposed system selects three sets of features based on previous studies: user-related, linguistics, and temporal propagation features. In the study, these features were selected through time series. Hidden Markov models (HMMs) were used to classify the tweets. First, two models were trained, one for rumors and the other for non-rumors over the time series. To predict the veracity of a new text, features

were extracted for each time series, and then, the features were passed to the two developed models. A probability score was computed by comparing the fitness of the passed features to the model. The probability scores that resulted from the comparison of both the true and false model were used to predict the rumor veracity. The system was evaluated using 209 rumors; it was able to correctly predict the veracity of 75% of them.

Deep learning was proposed to extract features from a large labeled data set from Twitter (Singh et al., 2019). Three different models were compared. The first model used four different classifiers. The second model used a deep learning model with a long-short-term memory (LSTM) network to analyze the tweet text, and the third model used a deep learning model with LSTM; however, this model utilized both the user characteristics and tweet text. When deep learning was used, an embedding vector was generated for each tweet text by listing all of the unique words for all of the tweets, and then, a pretrained look-up matrix (GloVe) was used to compute the embeddings for each word. An embedding vector was then used as input to the LSTM; it consisted of a 2-layer network model comprised of 100 neurons for each unit. The sentence length was 32-bits, and all of the short and long tweets were adjusted to fit that size. Then, each word was represented with a 200-length vector, which scaled down to 100-length after passing the first layer. After passing the second layer, the output is classified as rumor/non-rumor. In the third model, the user characteristics along with the tweet texts were used to predict the veracity. All of the information was used as input to the second layer of the LSTM. The

experimental study found that the deep learning model outperforms other classifiers. In addition, the tweet texts alone were sufficient to predict the veracity of the tweets.

In addition, in a recent study, Kumar et al. (Kumar et al., 2019) utilized particle swarm optimization (PSO) to improve the prediction process by improving the feature selection step. Five classifiers were used with the optimized feature selection to identify the most accurate classifier. First, thirteen features that are content features, pragmatic features (e.g., sentiment and name), and network features (e.g., account states and number of follows) were selected. To improve the performance of the selected classifiers, a PSO was used to select a subset of the selected features. A PSO is an iterative metaheuristic algorithm that searches the space for new solutions that are better than the selected best solution at any point in time until the optimal global solution for the whole swarm is found based on a determined number of iterations and fitness value. The results showed that using the PSO reached a 31% reduction in the feature selection, and it improved the accuracy for all of the classifiers except for SVM; the average improvement was 11.28%.

Devi et al. proposed a novel hybrid method that detected the source of rumors and then predicted the possibility of rumor spreading (Devi et al., 2020). The study was performed on Sydneysieg events. All of the tweets and their replies and retweets were collected. Conversation threads were classified as rumor/nonrumor. Furthermore, rumor threads were classified based on the support type in which each follower responded by either supporting the information of the source tweet or not. Both manual labeling and the proposed method detected the same threads to be sources of rumors. To predict the possibility of rumor spreading, a network of rumors with infected parts was developed. A breadth-first search (BFS) tree structure along with a spanning tree were created to calculate the probability for each node to be a source of rumors. The predicted nodes as the source of rumors matched the actual rumors.

Samule and Zaiane (Samuel and Zaiane, 2018) proposed the MedFact algorithm, which is based on evidence-based medicine (EBM) and trusted sources for medical information to judge the veracity (in terms of trustworthiness) of health information. EBM related health information is arranged in a hierarchical structure, starting from higher reliability sources to lower reliability sources. The proposed method consists of five steps. First, it extracts health phrases for which their veracity is not known using the TextRank algorithm. Then, an artificial neural network is used to classify the phrase as medically related or not. Automated information retrieval is used to find trusted articles that are related to the phrases; the articles are ranked based on their quality and usefulness. A set of trusted phrases is created for each article using word tokenization, part-of-speech (POS) tagging, and phrase segmentation. An agreement score, which is a semantic similarity score, is computed for each unknown phrase and the trusted phrases, using deep learning. Then, an aggregated agreement score is computed for the unknown phrase. Last, a veracity score is computed for the post in social media by using the aggregated agreement scores for all of the unknown phrases in that post. To measure the accuracy of the veracity score, the authors conducted a survey of nineteen users. Medical posts along with their veracity score and the top three trusted articles are displayed to the user. Then, they ask users about the usefulness of the resulting information; 67.54% of the users in the study gave positive feedback.

On the other hand, crowdsourcing technique has also been used to solve the big data veracity problem (Agarwal et al., 2016). *TagMe!* is a web-based application that was developed to extract tweets from Twitter and display them to users who classify the sentiments of the tweets as positive, negative, or neutral. Then, tagged data along with verified data that was downloaded from the sentiment analysis in Twitter (Rosenthal et al., 2019) are used to train a Bayesian predictor to predict the sentiment of the tweet. The results of the study showed that the proposed method was 89% accurate with an 11.42% average error.

Table 2 summarized existing work that addressed data veracity for social media applications.

Table 2: Data Veracity Studies for Social Media Applications.

Authors	Data set size	Data format	Approach	Performance
Wu et al. (Wu et al., 2015)	2601 false rumors each has 100 reports	Textual data (tweets) and quantities data (#flowers, #friends)	Predict rumors by comparing the similarities between propagation trees that account for the tweet's propagation structure and semantic features.	This approach is highly sensitive to the number of tweets and their propagation structure. After constructing the classifier model, the computational cost will be reduced significantly since it will be used for prediction only.
Ma et al. (Ma et al., 2015)	601 rumors and 537 non-rumors tweets	Textual data (tweets)	Predict rumors by presenting the dynamic series time structure model, which predicts rumors based on text features collected over a period of time.	Feature vectors are created for each time interval and for the changes between consecutive intervals. Then, the topic distribution and average sentiment score were computed for each post, which will increase both the computational and communication costs at this stage.
Giasemidis et al (Giasemidis et al., 2016)	72 rumors	Textual data and quantities data (#friends)	Predict the trustworthiness score by applying machine learning classifiers to predict rumors using the user's past behaviors.	Features are collected over 20 time series and in each interval from the time of rumor initiation. A high volume of data (high communication cost) is used to build the classifier model, which increased the computational cost.
Kwon et al. (Kwon et al., 2017)	130 rumors and non-rumors events	Textual and temporal data	Investigate the rumor characteristics over different time windows and determine the best feature selection by applying the random forest algorithm.	Data is collected over days starting from 3 days until 56 days. For each day, a set of features is collected. Thus, a higher communication cost is added on top of the computational cost of applying the random forest algorithm.
Olivieri et al (Olivieri et al., 2017)	20,000 reviews (positive and negative)	Textual data (movie reviews)	Predicted trustworthiness score for online input data using information about the domain of interest along with a knowledge base.	Depends on word meanings and more information related to the domain of interest. Thus, to obtain better prediction, the data volume must increase, which increases the communication cost. Computation cost relates to the time complexity if logistic regression algorithm which can be high.
Paryani et al. (Paryani et al., 2017)	63 millions tweets	Textual data	Measured text ambiguity of Twitter using an entropy-based model.	Topics are extracted from texts by counting word occurrences. Then, keyword probabilities are computed for all words that define the topic. This approach used a high volume of data, which will also impact the computational cost needed to compute the text ambiguity.
Vosoughi et al. (Vosoughi et al., 2017)	209 rumours	Textual and temporal data	Develop a system to predict the veracity of real-time rumors automatically using hidden Markov models.	This tool involves many computations, which increases its cost. However, the automation process can be a factor to manage this cost. Communication cost is high due to the collection of features of each time series.
Singh et al. (Singh et al., 2019)	5802 tweets	Textual data	Apply deep learning to extract features from large labeled data sets from Twitter.	Tweet texts alone improve the prediction, which reduces the data volume partially; however, deep learning requires a large data set to train the model. This requirement will increase both the computational and communication costs.

Authors	Data set size	Data format	Approach	Performance
Kumar et al. (Kumar et al., 2019)	14K tweets	Textual data	Utilize PSO to improve the prediction process by improving the feature selection step.	The PSO algorithm can be computationally costly; however, the stopping criteria can be determined to manage this cost. Thus, the computational time can be decreased.
Devi et al. (Devi et al., 2020)	71 conversation threads	Textul and temporal data	Detect the source of the rumor and the predicted rumor spreading using a hybrid method.	Tweet threads, including text and its replays, are collected, and a network of rumors is created to calculate the spreading probability for each node. This approach involves large volume of data and high computational cost.
Samule and Zaiane (Samuel and Zaiane, 2018)	3,958 questions and 2,260 tags	Textual medical data (phrases)	Measured veracity in terms of trustworthiness, applying evidencebased medicine algorithm and trusted sources.	Involves many algorithms, such as TextRank, neural networks, and deep learning algorithms, which increases both the computational and communication costs.
Agarwal et al. (Agarwal et al., 2016)	7226 tweets	Textual data	Develop a system based on a crowdsourcing technique to predict the tweets' sentiments.	This method is an automated approach that can improve the computational cost. However, it involves a human, which introduces some bias and increases the communication cost.

6. Web Applications

One of the largest sources for big data is web applications. Web data can be structured and unstructured; however, there is a move toward using structured data, called *linked data*, which provides consumers with a way to validate the information. This validation can be accomplished by linking the information to related sources. In addition, to make linked data more valuable, metadata that followed W3C PROV standards can be provided (Booth et al., 2004). Debattista et al. discussed linked data methodologies that can be applied to address data veracity (Debattista et al., 2015). They proposed two techniques that can be used when assessing data veracity in terms of specified quality metrics. The first technique is reservoir sampling, which measures dereferenceability and consistency in terms of disjoint classes. This random statistical technique is used to sample k items from equal distributions randomly. The number of selected items impact the computation time and the accuracy of the results. It is a fast solution when addressing big data because of its randomized nature. The other technique is the bloom filter, which is a common bit vector data structure technique that is used to query elements in a set, and it is used to measure the number of unique items. A hash function is created to map each item to its corresponding bits in the array filter. The accuracy of the results depends on the size of the bit vector data structure. One of the disadvantages of this technique is the production of a false positive. However, with some tuning, a Bloom filter can be useful for detecting duplications in data streams. Thus, there is a trade-off between the computation time and the precision.

Researchers have connected big data veracity with data uncertainty. Rubin and Lukoianova proposed measurements to assess and then manage the content and expression of textual big data (Rubin and Lukoianova, 2013). In terms of the content, the level of objectivity/subjectivity, deception/truthfulness, and credibility/implausibility (OTC) should be managed. However, to manage the expression,

sentence certainty should be evaluated by applying the analytical framework proposed by Rubin (Rubin, 2006). Structured textual data (verbal expression and content), that are collected via web applications using formats such as Microdata, RDFa or JSON-LD, was analyzed to identify a certainty level over a period of time using natural language processing based on four features: the certainty level; writer's and reporter's points of view; text focus, such as opinion, emotion, or fact; and time. Because OCT was proposed to be the main dimension of data veracity, subjectivity, deception, and implausibility (SDI) were the quantitative measurements used to validate the information. Thus, measuring the level of SDI provides information about SDI, which reduces the uncertainty. Then, the veracity index can be calculated; the index is the average of the SDI levels. Textual context with a high SDI level must be processed and cleaned and, if possible, not used in the decision-making process. Existing tools were used to measure SDI. Subjectivity and opinion detection can be measured by an automated classification tool that analyzes text in terms of sentiments or opinion (Hirst, 2007). There are many deception detection tools for natural language processing that depend on linguistics and machine learning to remove human bias and speed up the detection process. On the other hand, tools that analyze personal opinions, evaluations, and recommendations can play a major role in measuring credibility.

Another study computed the data veracity for multisource structured data in terms of a confidence score and trustworthiness. Truth discovery methods, specifically a single truth discovery method, are used to compute the confidence score and trustworthiness for each data value and each data source for multi-sources data (BertiÉquille, 2015). However, there is an issue related to the unavailability and bias of ground truth data, where veracity is checked and labeled manually to be used later in the evaluation method, especially in the case of big data. To solve this issue, which involves discovery method selection and truth data bias, Berti proposed ensemble truth discovery methods to improve the resulting accuracy (Berti-Équille, 2015). Two different ensemble methods were defined: uniform weight and adjusted weight for the combined methods. The adjusted weight approach assigns a weight to each method that reflects its accuracy. The experiment using claims about researchers' affiliation provided by different users/sources. The results using real-world data and synthetic data showed that ensemble methods with both uniform and adjusted weights have better accuracy than most accurate single methods. Additionally, the ensemble methods gave, on average, the best-quality performance when computing data veracity with small ground truth datasets.

McArdle and Kitchin focused on the veracity of open databases for real-time urban data (McArdle and Kitchin, 2016). However, because the need for open databases increases, there is a risk of impacting the data veracity, which would lead to poor decisions and "buggy" applications. Two applications that depend on different sources of real-time urban data were studied to detect and repair the data problems. The investigators found that although the standards and guidelines and core quality metrics were defined for urban data (Guptill and Morrison, 2013; Turner, 2004), the data producer provided data without using quality metrics and with no information about the data quality. The authors suggested that the data producer should share the document issues related to metadata and address how consumers can address these issues in such a way that they can enjoy the best utilization of the data for a specified purpose. They also proposed using crowdsourcing mechanisms to record user observations and fixes, to improve the data veracity.

In addition, the veracity of the news has been assessed using emotional analysis (Tarmizi et al., 2019). Emotional states are detected, and a weight is calculated for each state; the weights are used to differentiate between legitimate and fake news. This study focused on a set of emotions, such as anger, fear, sadness, and joy. First, words were identified from news text and labeled with the corresponding emotion using an emotion lexicon called EmoLax. This lexicon uses the frequency of the sentence

components (e.g., nouns and verbs). Then, a weight was computed for each emotion state. To validate the presented approach, two datasets were used: The Onion and Star, where The Onion is an online news that consists of fake news, and Star is a Malaysian news organization. The authors found that even legitimate news consists of negative emotional states, and thus, such states cannot be used as a sign for fake news detection. However, the authors also found that fake news has higher emotional weights, which indicates the use of exaggerated emotional states to affect the reader's beliefs.

Data veracity has been improved using different techniques. A cryptography technique has been used to ensure the data veracity. Traditional techniques introduce overhead, which can be inefficient when applied to big data. A study by Kepner et al. proposed a technique, known as the computing on masked data (CMD) system, to improve the data veracity by performing computations on masked data in which only an authorized recipient can unmask the data (Kepner et al., 2014). The CMD depends on the associative array, which includes features from both sparse matrices and database stored records. Associative arrays show complex relationships between entities and are represented as a sparse matrix or graph. Thus, the system performs both matrix and graph algorithms due to its structure. The CMD was evaluated using two examples: DNA sequence matching algorithms and database operations on social media. It reduced the overhead while revealing less information about the actual data, which preserves the data privacy and, as a result, its veracity.

Blockchain has also been proposed as a new approach to improve data veracity (Wibowo and Sandikapura, 2019). However, blockchain cannot ensure the veracity of the data; it validates the input devices. Thus, it allows a user to transfer data securely using the concept of consensus without third parties. Token-based crowdsourcing and source identification have been proposed to improve the data veracity. This approach provides tokens as rewards for sharing and validation of data. Wibowo and Sumari (Wibowo and Sumari, 2020) described the HARA tool, which is used as a token-based crowdsourcing technique, to improve the veracity of the input data. It is a decentralized ecosystem for data exchange. It consists of data providers, data buyers, and data qualifiers; each has its own roles for which they receive tokens. Data providers, such as organizations or individuals, provide their data to the system to be checked for quality and then receive tokens plus some other benefits provided by HARA for sharing their data. As a response, data qualifiers verify the data and receive tokens. Then, data buyers (aka enterprises) access data through the systems and will obtain tokens through sharing the analyzed data and results. Another approach is source identification for data origin, derivation, and ownership to improve its trustworthiness. However, this approach suffers from privacy risks due to sharing the source identity. Therefore, a decentralized identifier (DID) was proposed to solve this issue. The DID contains cryptographic data that authenticates the entity of the specified DID since the DID is under the control of the subject. For more details on generating DID, check (Reed et al., 2018).

Failure mode and effect analysis (FMEA) is a different approach that was proposed to improve both the veracity and validity of the data (DVV-FMEA) (Herrera et al., 2019). It lists the system components and the relationships between them as well as the failure modes and the reason for each failure. Then, FMEA is used to score each failure based on its severity level. Detection tools are used to determine the detectability score for each failure. The last step is to prioritize the risk by computing the risk priority number (RPN), which is obtained by multiplying the severity, occurrence, and detectability scores. Then, a product or service is improved based on its order. DVV-FMEA is based on FMEA; however, it has been modified to improve the data validity and veracity. For example, to identify the causes of the failures, reasons related to the data, such as data integration and data ambiguity, are considered. Moreover, FMEA considers the impact of data integrity, availability, and consistency. The effectiveness of

DVV-FMEA was demonstrated using a production testing database of electronic devices; it was found to improve both the data veracity and the data validity.

A summary of existing research for data veracity in the context of web applications is shown in Table 3.

Table 3: Data Veracity Studies for Web Applications.

Authors	Data set size	Data format	Approach	Performance
Debattista et al. (Debattista et al., 2015)	Not mentioned	Structured linked web data	Measured veracity in terms of dereferenceability, consistency, and items' uniqueness using reservoir sampling and bloom filter.	Improves the computational cost of quality metrics when addressing big data; parameters tuning for the bloom filter can improve the computational cost even more at the cost of its precision.
Robin and Lukoianova (Rubin and Lukoianova, 2013)	Not mentioned	Textual data (verbal expression and content)	Measured content and expression uncertainty using machine learning and natural language processing, to measure the subjectivity, deception, and implausibility (SDI) as quantitative measurements.	Texts are analyzed, which requires a large amount of time; however, this approach depends on existing tools to automate the process, which can lower the computational cost. It also involves high communication cost since it depends on text analysis.
Berti-Équille (Berti-Équille, 2015)	43,245 claims of author name	Textual data (Structured claims)	Measured confidence score and trustworthiness, applying ensemble truth discovery methods for multisources data	Computes the confidence score and trustworthiness depending on the ground truth table, which is a manual process and involves human bias. Thus, there is a high computational cost. Scores are computed for each data value and data source which indicates a high communication cost.
McArdle and Kitchin (McArdle and Kitchin, 2016)	Not mentioned	Urban data	Improve the veracity through document sharing along with issues related to metadata and how the user can address these issues; a crowdsourcing approach is followed to record the user's observations and fixes.	Crowdsourcing suffers from many issues related to data volume, data reliability, and data variety, in addition to the privacy of the data providers. Thus, high computational and communication costs.
Tarmizi et al. (Tarmizi et al., 2019)	3,235 news records	Textual data (emotions)	Computed weight for each emotional state using emotion lexicon to identify words, label them, and then compute the frequency of the sentences component.	Text analysis increased the computational cost. However, since it focuses on emotions only, the number of reviewed and classified words is limited, which improves the cost in terms of computation and communication.
Kepner et al. (Kepner et al., 2014)	50,000 tweets	Textual data	Improve the data veracity by performing computations on masked data in which only an authorized recipient can unmask the data.	It reduces the overhead which was introduced by a previous work. However, mask and unmask data can add computational cost. Communication cost depends on the size of data that will masked/unmasked.

Authors	Data set size	Data format	Approach	Performance
Wibowo et al. (Wibowo and Sandikapura, 2019)	Not mentioned	Quantities data via Sensors	Validate the input devices using blockchain by allowing the user to transfer data securely using the concept of consensus without third parties, utilizing token-based crowdsourcing and source identification.	Blockchain involves a high volume of communication that can increase the computational time since it depends on the contributions of data providers and users.
Herrera et al. (Herrera et al., 2019)	68,168 devices	IDs such as cell number, temperature, tester ID, operator ID, etc	Improve the veracity of the data using failure mode and effect analysis.	Many computations and different components are involved to detect and prioritize failures, which can impact both the computational and communication costs.

7. Internet of Things Applications

The internet of things (IoT) generates big data through the use of standalone or embedded sensors. However, the collected data suffers from many issues, including inconsistency and incompleteness. Thus, the veracity of these data is very crucial, and studies have focused on studying it in terms of data uncertainty and the reputations of data sources.

Data uncertainties in IoT applications have been widely studied. Data streams that are read by radio frequency identification (RFID) are unreliable. To fix data issues, middleware systems that connect readers to applications use a "smoothing filter," which is a sliding window that inserts values for lost ones. The window size has a great impact when cleaning a data stream. In addition, there is a need to have a fixed size for each application. Jeffery et al. proposed a Statistical sMoothing for Unreliable RFid data (SMURF) that determines window size automatically for a period of time for each application (Jeffery et al., 2006). SMURF is based on a samplingbased approach that randomly samples data streams that are then used to derive adaptive smoothing techniques. SMURF was evaluated using both synthetic and real RFID data streams and was shown to produce more reliable data streams.

To capture uncertainty in high-volume streams coming from sensors, Diao et al. presented probabilistic modeling and inferencing to present time and space efficiently (Diao et al., 2009). Data is modeled as a continuous random variable using graphical modeling. The observed data used probabilistic inference to transform data into a format (e.g., tuples) that is suitable for more processing. Then, data uncertainty is described using probability density functions that are inserted for each tuple. To handle the speed of data streams efficiently, their approach used particle filtering, which is a sampling-based inference. The proposed optimizations improved particle filtering by processing 1000 readings per seconds for 20K objects compared to 0.1 readings per second for 20 objects.

Rodríguez et al. proposed an approach to manage data uncertainty for sensor data following the quality principles of data (Rodríguez and Servigne, 2013). The proposed approach focused on the quantification and communication of data quality through the specification of data quality sources, the estimation of sensor data quality, and the management of data quality information. Data quality sources were related to the quality of data and their metadata. The metadata consists of sensor data such as status, battery level, storage level, and sensor coordinates and manages the data quality properties. The

data quality is then evaluated using and communicated with users via visualization, audio signals, and/or reports. Reports are generated based on the behavior of the system or on user preference.

Chen and Jiang proposed a parallel algorithm that combines a MapReduce model with traditional preprocessing methods to clean large datasets with missing data (Chen and Jiang, 2014). The algorithm implemented Java and Hadoop, which, using many nodes, process applications with large datasets in parallel. They experimentally tested the performance of the proposed algorithm using a KDDCUP98 data set, which was divided into three sets: DS1, consisting of 19,177 records, DS2, consisting of 191,779 records, and DS3, consisting of 1,917,790 records. The experiment was conducted using nine nodes, and they found that the time required to fix missing data was improved as the number of nodes increased. The efficiency of the algorithm was then studied with a larger dataset and was found to be more efficient.

Zhang et al. proposed a parallel matrix-based method based on rough set theory to compute an approximation

of missing data (Zhang et al., 2014). MapReduce was used to implement a parallel method to handle large amounts of data. The proposed method was evaluated on two datasets from the University of California at Irvine with a total sample of 8,749 with a large number of attributes. Different strategies were implemented: parallel strategy based on MapReduce, incremental parallel strategy based on MapReduce, and incremental parallel strategy using a sparse matrix based on MapReduce. The results found that the first strategy was efficient when processing large datasets; however, the incremental process demonstrated better performance.

Sanyal and Zhang improved the veracity of data in sensor data that are collected using device to device communication through a data aggregation scheme for fixing data uncertainty (Sanyal and Zhang, 2018). The proposed method is an iterative process that tracks the optimal dominant subspace. The optimal subspace is then used to estimate the data matrix of the uncertain data. The presented approach can estimate sensor data in the presence of missing data, outliers, and noise compared to a principal component analysis (PCA).

Real-time big data that are collected via sensors in smart electronic or transportation grids can be unavailable due to network or consumer limitations. Thus, only part of the data is available. Different methods have been proposed to improve the data veracity by predicting the amount of missing data (Deshpande et al., 2004; Kreindler and Lumsden, 2006; Razzaque et al., 2013); however, one study predicted missing data using partially available data based on the dependencies between the data of each time period (Aman et al., 2014). Thus, the developed influence model (IM) consisted of two stages. In the first stage, the learning phase, data that are collected from smart sensors are correlated and divided based on the time series. In the second stage, a prediction model was built to discover missing data. The model accuracy was evaluated using the data of Los Angeles electricity consumption, which was collected by smart meters, in addition to weather data that was collected by the Los Angeles NOAA's USC station. The model used 8-hour intervals. The baseline method, which is an auto-regressive tree (ART), performed well for 6-hours out of the 8-hour intervals, and then, it became inefficient.

IM outperformed ART: it reduced the mean absolute percentage error by 10%.

Amini and Chang (Amini et al., 2016) proposed a modified version of TOPSIS to calculate the ranking distribution of different sensors. TOPSIS is a multi-attribute decision-making (MADM) method that ranks different sources; it was first introduced by Hwang and Yoon (Tzeng and Huang, 2011). The main concept of TOPSIS consists of a "negative ideal solution (NIS)" and a "positive ideal solution (PIS)", where PIS is the best score of an attribute. The best option is the one with the closest score to the PIS. The authors conducted 100 simulation runs over an interval-based score; uncertain intervals

and scores were provided by decision makers. Then, the modified-TOPSIS computed the closeness coefficients of each sensor over 100 runs; the closeness coefficient scores reflected the trustworthiness of the data with respect to the specified quality attributes.

Mobile cloud computing (MCC) is another promising paradigm that provides data storage and processes to users. Large amounts of data, such as an individual's location, vital signs, and health records, are collected through MCC. These data are then used for decision-making. Thus, it is important to maintain the veracity of the data. In (Lin et al., 2017; Lin et al., 2015), the authors proposed a new category-based context-aware and recommendation incentive reputation mechanism (CCRM) that addresses data worthiness to improve the data veracity, with a reputation mechanism that is used as a defense against any internal attacks in MCC. In this mechanism, the data are classified based on the required security level, a higher security level leads to a stricter reputation mechanism. The CCRM combines different reputation evaluations: a data category based on the security level, context-aware technologies, and a Vickrey-Clark-Groves (VCG) mechanism. The experimental study found that the CCRM performs better than the recommendation and privacy preserving-based cross-layer reputation mechanism (RP-CRM), the anonymous reputation and trust in participatory sensing (ARTSense), and harmony mechanisms, especially with collusion attacks and bad mouthing attacks.

Recently, trust-based data collection was introduced to control the veracity of data collected via sensors by giving each data collector different trust values based on their performance during the data collection stage. In this approach, higher values are given to trusted parties. Jiang et al. (Jiang et al., 2020) used an unmanned aerial vehicle (UAV) as a baseline to evaluate other collecting nodes. Thus, in this approach, sensors are evaluated by comparing the collected data to the baseline. If the data is within a predefined error range, then the sensor is considered to be trusted and its trust value is increased.

The system will filter sensors that have a low trust value. To evaluate this approach, a simulation experiment was conducted to create a network area with a set of sensors using SCADA characteristics, and it was repeated for twenty different scenarios. The results showed that using UAV helped to evaluate trust sensors with an accuracy of 778% after 100 rounds, and the accuracy increased as the round number increased.

The other work by Li et al. (Li et al., 2020) proposed a trust-based data collection system that depends on UVAs and sensors embedded in vehicles to collect the data. The proposed system aims at improving the security of the data by identifying the trusted vehicles. A data center collects data from trusted vehicles via UVAs. Data in the trusted vehicles are collected from other untrusted vehicles, called ordinary vehicles. Trusted vehicles are selected based on their trajectories. The authors hypothesized that vehicles with fixed parking places are more trusted than other vehicles with no fixed parking spaces. Thus, historical data for a vehicle is used to compute the trustworthiness ratio, which is related to the frequency of parking vehicles in fixed places; a higher ratio indicates more trusted vehicles. Since vehicles cannot cover all possible locations in a city, static sensors are used also to collect data. These sensors are placed in locations where no trusted vehicles are close enough to a data center. Then, UVAs are guided to collect data from these trusted vehicles and static stations at specific times. The experimental study performed on T-driver datasets in Beijing city showed that the proposed method utilizes more vehicles for data collection since it allows data transformation from ordinary vehicles to trusted vehicles and that the number of malicious attacks decreases significantly.

Ren et al. proposed a trust-based minimum cost quality aware (TMCQA) data collection scheme (Ren et al., 2020) to predict the most trusted source of data, called a data reporter. Not all data reporters are trusted since malicious attacks can impact the veracity of the collected data. Thus, the proposed scheme uses a machine learning algorithm to predict trusted reporters based on their historical

information over a period of time. Then, a trust value for the current time can be predicted. A reporter with a high trust value was selected in the study. For more accurate computation, the trust value is assigned different weights depending on the period. If a trust value is computed for the time period that is closer to the current time, a higher weight will be assigned. The performance of the proposed method was compared to a contribution-based with no trust value scheme (CNTVS), random data reporter selection scheme (RDRSS), and a trust-based with no time decay scheme (TNTDS). The authors found that TMCSQ improved the quality of service even when the reported behavior changed (e.g., from an attacker to a perfect reporter) over time.

The following table 4 summarized existing work that focused on assessing, predicting, and improving veracity of data in IoT applications.

Table 4: Data Veracity Studies for IoT Applications

Authors	**Data set size**	**Data format**	**Approach**	**Performance**
Jeffery et al. (Jeffery et al., 2006)	Data for 5000 epochs	RFID data stream	Random sampling filter is used to determine windows size automatically that is used for a period of time for each application	Expected computational and communication costs will be high since this approach deals with data stream. However, the determined window size is fixed and can be used for long period which can reduce computational time/cost.
Diano et al. (Diao et al., 2009)	1000 readings per seconds for 20K objects	Data streams	Present a probabilistic modeling and inference to present time and space efficiently	Communication and computational costs are high for data streams; however, sampling-based inference was used to speed up the process which improve the reading speed, thus improving the costs.
Chen and Jiang (Chen and Jiang, 2014)	2128746 records	Sensors data	Developed parallel algorithm that combines Mapreduce model with the traditional pre-processing methods to clean large dataset with missing data	Computational time will be reduced for the development of parallel algorithm. However, the communications costs will be high since it focused on large datasets.
Zhang et al. (Zhang et al., 2014)	8749 samples	Categorical data	Developed a parallel method using Mapreduce to handle incompletion of large amount of data.	Computational time will be reduced due to the use of parallel algorithms. However, the communications costs will be high since it handle fast amount of data
Sanyal and Zhang (Sanyal and Zhang, 2018)	Data with 10 subspace	Sensors data	Proposed data aggregation scheme that iteratively tracks the optimal dominant subspace to estimate data matrix for uncertain data	Computational and communication costs are slightly high in D2D communications. In addition to the cost of iterative approach.
Aman et al. (Aman et al., 2014)	50,000 meters	Sensor data	Improve the data veracity by developing an influence model that predicts the values of the missing data.	Data and their dependencies are collected from sensors for each time period. Thus, the expected computational and communication costs will be high.

Authors	Data set size	Data format	Approach	Performance
Amini and Chang (Amini et al., 2016)	100 iterations to generate 100 ranks	Sensors data	Calculated ranking distributions of different sensors by computing closeness coefficient, which represents trustworthiness given the uncertainty intervals and score.	Involves many calculations and ordering for each senors. The whole calculations must be repeated when any new sensor is used. In addition, this method suffers from human bias since uncertainty scores is provided by decision makers which increase both computational and communication costs.
Lin et al. (Lin et al., 2017; Lin et al., 2015)	100 mobile clients, 10 routers, 10 services providers	Data collected via Mobile cloud computing	Improve the data veracity through category-based context-awareness and a recommendation incentive reputation mechanism to defend any internal attacks.	Reputation mechanism requires communicating with the user's neighbors. Thus, there is a high computational and communication costs. However, these costs can be improved when the number of neighbors is minimized.
Jiang et al. (Jiang et al., 2020)	20 different networks scenarios, and 36 to 40 sensors	UAV data	Measured trust value for each data collector using UAV as a baseline to evaluate the collecting nodes.	Data collection depends on the network speed to transfer the data, which might slow the process of identifying the trusted data collector (high computational costs). Trust values is computed and evaluated for each sensor which increased communication cost.
Li et al. (Li et al., 2020)	18 trusted vehicles, 4 UVA collector regions, and 5 static sensors	UAV data and Vehicles data (locations)	Measured trustworthiness ratio for each vehicle using UVA and trusted vehicles to collect data.	Identifies trusted vehicles. Then, UAV will communicate with only selected vehicles to collect the data. This approach reduces the communication cost by limiting the number of data collectors. However, data data volume depends on the number of vehicles; high number leads to high communication cost.
Ren et al. (Ren et al., 2020)	3 data reporter and 80 rounds	Sensor data	Predict trust value of current time to predict the trusted reporter using historical information over a period of time.	The computational and communication cost for this approach will increase as the number of reporters and time windows increase. However, a cost is paid once since after building the model, it will be used to predict the trust values.

8. Challenges and Future Directions

Data veracity is an essential dimension of big data. Some researchers have studied in terms of data quality dimensions, which are considered by some to be data access authorization, uncertainty, incompleteness, trustworthiness, and more. Before more studies are performed in regard to data veracity, a standard definition is needed since veracity definitions have been defined differently based on the related field. For example, in social media applications, veracity has been measured in terms of data trustworthiness and they focused on the identification of rumors and fake news. On the other hand, security specialists have studied data veracity in terms of access authorization for web data. These are totally different perspectives; thus, we believe that data veracity should have one general standard definition that covers the related aspects and should be used to direct research in the area.

Researchers in many computer science fields have contributed to assess the data veracity. We have surveyed existing work in this regard. Most of the proposed approaches for veracity assessment have involved a manual or semi-manual process, which represents a challenge to the assessment of the data veracity, especially in the area of big data.

Most of the studies in the literature utilized machine learning (ML) algorithms for data veracity prediction. Some studies predicted a value that represents trustworthiness of the data collector, and most used ML algorithms to detect rumors and their sources in social media applications. There was some work pursued with the aim of automating the process of prediction. Most of the existing work used Twitter textual and temporal data, which might be related to its availability.

Data collection is time consuming and introduces a challenge for most fields due to data unavailability or data volume. In addition, predicting data veracity suffers from many issues that arise from data uncertainty, data provenance, and noisy data (L'Heureux et al., 2017). Data are uncertain due to the veracity of the data collection sources. For example, social media data is produced by humans, and sensor data are collected by different sensors; both can introduce noise to the data. Data provenance relates to data tracking and recording, which involves a high volume of data and, as a result, increases the computational cost. Last, noise in data that are introduced by the veracity of data sources can confuse ML algorithms when attempting to make the correct prediction. In addition, most of the work on data veracity involved text analysis, and thus, it suffers from issues related to this field, such as text language (slang vs. standard), mixed languages, and grammar and spelling errors (Shahnawaz and Astya, 2017).

Some of the group of work presented focused on improving data veracity by utilizing security methods such as mask data to authorize access, applying blockchain to transfer data securely, or utilizing crowdsourcing to fix the data issues. In addition, ML algorithms were used to predict missing values.

The presented methods improved the data veracity from different perspectives, and each has its own challenges. One issue is related to the high volume of data due to the large number of data sources in crowdsourcing. In addition, this approach depends on users' contributions, which introduces data uncertainties and incompleteness (Jagadish et al., 2014). Crowdsourcing suffers from other issues, such as data reliability and user privacy. These issues can degrade the data veracity (Srivastava and Mostafavi, 2018). Another challenge relates to blockchain; it suffers from different types of attacks which can degrade the veracity of data. In addition, due to the communication overhead that is introduced from consensus mechanism, the performance of blockchain framework is poor (Gao et al., 2018).

Comparing the presented techniques based on their application domains, we found that the most useful technique for addressing data veracity in social media applications was the use of machine learning algorithms that analyze and classify text, as well as a reduction of feature selections, which in response will improve both the computational and the communication costs. On the other hand, for web applications, crowdsourcing techniques provide more information that can be used to assess and improve veracity. For IoT applications, identifying trusted sources based on their reputations helps to improve the veracity of data collection. The use of trusted sources determined by the use of UAV and embedded sensors in mobile phones and vehicles has been found to be the best presented method to validate data collected via sensors.

The presented techniques can be used across application domains since the underlying problem is the veracity of big data. For example, machine learning techniques have been widely used in social media applications to predict rumors; however, they can be adapted by web or IoT cities applications if the goal is to predict false data sources, as proposed by Robin and Lukoianova (Rubin and

Lukoianova, 2013) for web data and Ren et al. (Ren et al., 2020) for IoT applications. In addition, the crowdsourcing technique depends on the contributions of different parties to evaluate data sources. Such an approach is applicable for social media applications (Agarwal et al., 2016), as well as for web applications (McArdle and Kitchin, 2016) and IoT applications (Lin et al., 2017; Lin et al., 2015).

Most of the work presented herein introduced high computational and communication cost since they depended on whole data sets and involved many computations/processes to gain insight into the data veracity. It is not clear in the existing studies whether these data are performed on the fly or are saved for later evaluation. Thus, there is no clear insight about the storage space. However, if we assume that the computations were saved, then we expect the need to have a large storage space since these calculations were performed on all of the objects in the data sets.

Much work is still needed in this area. Data are collected from different sources; there is currently no work in the literature that compares the data collection methods and their impact on the veracity, as well as a recommendation for the best method for data collection with respect to the veracity. In addition, a veracity manual assessment method introduces a burden to the process, and thus, we propose to motivate the development of a tool to automate the process. Most of the work presented used social media data; more comprehensive work is needed to use data from other domains to identify the best method to improve the data veracity to its maximum potential. One of the issues of big data is the volume. Thus, reducing the data dimensions through feature selection could be one way to reduce the communication cost. More work is needed on aspects from different domains, such as slicing in software engineering, which was introduced to speed up software debugging (Reps et al., 1994; De Lucia, 2001); the aspect of slicing can be adapted to slice data when studying data veracity to reduce the data volume. There is no study on introducing distributed systems and the benefit that they could offer toward improving the computational cost when addressing data veracity.

9. Conclusions

Data veracity impacts the decision-making process in different domains, such as social media, the web, and IoT. Evaluating the veracity of the data is no longer optional; it is one of the first steps that should be taken during data preprocessing to guarantee the quality of the results. Data should not be considered before validation. Our review summarized the existing work. A moderate amount of work was proposed to tackle the issue of data veracity; however, most of the studies focused on veracity aspects that are related to social media applications. We argue the effectiveness of these methods with data coming from different domains. In addition, data volume and veracity impact the computational and communication cost of the presented approaches; therefore, ideas from different fields, such as software engineering and parallel computing, can be very advantageous. More attention should be taken to improve data veracity. We also recommend investigating real cases from related industries to determine the needs of businesses as it relates to ensuring the veracity of data in real projects.

10. References

Agarwal, B., Ravikumar, A., and Saha, S., 2016. A Novel Approach to Big Data Veracity Using Crowdsourcing Techniques and Bayesian Predictors. In *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1020-1023.

Aman, S., Chelmis, C., and Prasanna, V., 2014. Addressing data veracity in big data applications. In *2014 IEEE International Conference on Big Data (Big Data)*, pages 1-3. IEEE.

Amini, M., Chang, S., and Malmir, B., 2016. A fuzzy MADM method for uncertain attributes using ranking distribution. In *Proceedings of the industrial and systems engineering research conference*.

Batini, C., Cappiello, C., Francalanci, C., and Maurino, A., 2009. Methodologies for data quality assessment and improvement. *ACM computing surveys (CSUR)*, 41(3):1-52.

Batista, A. F., da Silva, D. L., and Correa, P. L., 2017. Enabling Data Legitimacy in Data-Driven Projects. In *2017 International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC)*, pages 50-54. IEEE.

Berti-Equille, L. and Borge-Holthoefer, J., 2015. *Veracity of Data: From Truth Discovery Computation Algorithms to Models of Misinformation Dynamics*. Morgan and Claypool.

Berti-Équille, L., 2015. Data veracity estimation with ensembling truth discovery methods. In *2015 IEEE International Conference on Big Data (Big Data)*, pages 2628-2636.

Booth, D., Haas, H., McCabe, F., Newcomer, E., Champion, M., Ferris, C., and Orchard, D., 2004. W3C working group note 11: Web Services architecture. *World Wide Web Consortium (W3C)*.

Chen, F. and Jiang, L., 2014. A parallel algorithm for datacleansing in incomplete information systems using mapreduce. In *2014 Tenth International Conference on Computational Intelligence and Security*, pages 273-277. IEEE.

De Lucia, A., 2001. Program slicing: Methods and applications. In *Proceedings First IEEE International Workshop on Source Code Analysis and Manipulation*, pages 142-149. IEEE.

Debattista, J., Lange, C., Scerri, S., and Auer, S., 2015. Linked'Big'Data: towards a manifold increase in big data value and veracity. In *2015 IEEE/ACM 2nd International Symposium on Big Data Computing (BDC)*, pages 92-98. IEEE.

Deshpande, A., Guestrin, C., Madden, S. R., Hellerstein, J. M., and Hong, W., 2004. Model-driven data acquisition in sensor networks. In *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*, pages 588-599.

Devi, P. S., Karthika, S., Venugopal, P., and Geetha, R., 2020. Veracity Analysis and Prediction in Social Big Data. In *Information and Communication Technology for Sustainable Development*, pages 289-298. Springer.

Diao, Y., Li, B., Liu, A., Peng, L., Sutton, C., Tran, T., and Zink, M., 2009. Capturing data uncertainty in high-volume stream processing. *arXiv preprint arXiv:0909.1777*.

Elloumi, O., Block, T. D., and Samovich, N., 2019. Market Drivers and High Level Architecture for IoT-enabled Data Market places. Technical report.

Gao, W., Hatcher, W. G., and Yu, W., 2018. A Survey of Blockchain: Techniques, Applications, and Challenges. In *2018 27th International Conference on Computer Communication and Networks (ICCCN)*, pages 1-11.

García Holgado, A., Marcos Pablos, S., García Peñalvo, F. J. et al., 2020. Guidelines for performing Systematic Research Projects Reviews. *International Journal of Interactive Multimedia and Artificial Intelligence*, 6(2):9.

Giasemidis, G., Singleton, C., Agrafiotis, I., Nurse, J. R., Pilgrim, A., Willis, C., and Greetham, D. V., 2016. Determining the veracity of rumours on Twitter. In *International Conference on Social Informatics*, pages 185-205. Springer.

Guptill, S. C. and Morrison, J. L., 2013. *Elements of spatial data quality*. Elsevier.

Herrera, A. E. H., Walshaw, C., Bailey, C., and Yin, C., 2019. Failure Mode Effect Analysis for Improving Data Veracity and Validity. In *2019 International Conference on Computing, Electronics Communications Engineering (iCCECE)*.

Hirst, G., 2007. *Views of Text Meaning in Computational Linguistics: Past, Present, and Future*. na.

Jagadish, H. V., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J. M., Ramakrishnan, R., and Shahabi, C., 2014. Big data and its technical challenges. *Communications of the ACM*, 57(7):86-94.

Jamil, N. B. C. E. ., Ishak, I. B., Sidi, F., Affendey, L. S., and Mamat, A., 2015. A Systematic Review on the Profiling of Digital News Portal for Big Data Veracity. *Procedia Computer Science*, 72:390-397. ISSN 1877-0509.

Jeffery, S. R., Garofalakis, M., and Franklin, M. J., 2006. Adaptive cleaning for RFID data streams. In *Vldb*, volume 6, pages 163-174. Citeseer.

Jiang, B., Huang, G., Wang, T., Gui, J., and Zhu, X., 2020. Trust based energy efficient data collection with unmanned aerial vehicle in edge network. *Transactions on Emerging Telecommunications Technologies*, page e3942.

Kaisler, S., Armour, F., Espinosa, J. A., and Money, W., 2013. Big data: Issues and challenges moving forward. In *2013 46th Hawaii International Conference on System Sciences*, pages 995-1004. IEEE.

Kepner, J., Gadepally, V., Michaleas, P., Schear, N., Varia, M., Yerukhimovich, A., and Cunningham, R. K., 2014. Computing on masked data: a high performance method for improving big data veracity. In *2014 IEEE High Performance Extreme Computing Conference (HPEC)*, pages 1-6. IEEE.

Kitchenham, B. and Charters, S., 2007. Guidelines for performing systematic literature reviews in software engineering.

Klein, A. and Lehner, W., 2009. Representing data quality in sensor data streaming environments. *Journal of Data and Information Quality (JDIQ)*, 1(2):1-28.

Kreindler, D. M. and Lumsden, C. J., 2006. The effects of the irregular sample and missing data in time series analysis. *Nonlinear dynamics, psychology, and life sciences*.

Kumar, A., Sangwan, S. R., and Nayyar, A., 2019. Rumour veracity detection on twitter using particle swarm optimized shallow classifiers. *Multimedia Tools and Applications*, 78(17):24083-24101.

Kwon, S., Cha, M., and Jung, K., 2017. Rumor detection over varying time windows. *PloS one*, 12(1).

Li, T., Liu, W., Wang, T., Ming, Z., Li, X., and Ma, M., 2020. Trust data collections via vehicles joint with unmanned aerial vehicles in the smart Internet of Things. *Transactions on Emerging Telecommunications Technologies*, page e3956.

Lin, H., Hu, J., Liu, J., Xu, L., and Wu, Y., 2015. A Context Aware Reputation Mechanism for Enhancing Big Data Veracity in Mobile Cloud Computing. In *2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing*, pages 2049-2054.

Lin, H., Hu, J., Tian, Y., Yang, L., and Xu, L., 2017. Toward better data veracity in mobile cloud computing: A context-aware and incentive-based reputation mechanism. *Information Sciences*, 387:238-253.

Liu, X., Tamminen, S., Su, X., Siirtola, P., Röning, J., Riekki, J., Kiljander, J., and Soininen, J.-P., 2018. Enhancing Veracity of IoT Generated Big Data in Decision Making. In *2018 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, pages 149-154. IEEE.

Lozano, M. G., Brynielsson, J., Franke, U., Rosell, M., Tjörnhammar, E., Varga, S., and Vlassov, V., 2020. Veracity assessment of online data. *Decision Support Systems*, 129:113132.

Lozano, M. G., Franke, U., Rosell, M., and Vlassov, V., 2015. Towards automatic veracity assessment of open source information. In *2015 IEEE International Congress on Big Data*, pages 199-206. IEEE.

L'Heureux, A., Grolinger, K., Elyamany, H. F., and Capretz, M. A. M., 2017. Machine Learning With Big Data: Challenges and Approaches. *IEEE Access*, 5:7776-7797.

Ma, J., Gao, W., Wei, Z., Lu, Y., and Wong, K.-F., 2015. Detect rumors using time series of social context information on microblogging websites. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 1751-1754.

McArdle, G. and Kitchin, R., 2016. Improving the veracity of open and real-time urban data. *Built Environment*, 42(3):457-473.

Moyne, J. and Iskandar, J., 2017. Big data analytics for smart manufacturing: Case studies in semiconductor manufacturing. *Processes*, 5(3):39.

Oguz, D., Ergenc, B., Yin, S., Dikenelli, O., and Hameurlain, A., 2015. Federated query processing on linkedˇ data: a qualitative survey and open challenges.

Olivieri, A. C., Shabani, S., Sokhn, M., and Cudré-Mauroux, P., 2017. Assessing data veracity through domain specific knowledge base inspection. In *2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pages 291-296.

Paryani, J., TK, A. K., and George, K., 2017. Entropy-Based Model for Estimating Veracity of Topics from Tweets. In *International Conference on Computational Collective Intelligence*, pages 417-427. Springer.

Patgiri, R. and Ahmed, A., 2016. Big data: The v's of the game changer paradigm. In *2016 IEEE 18th International Conference on High Performance Computing and Communications; IEEE 14th International Conference on Smart City; IEEE 2nd International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, pages 17-24. IEEE.

Ramachandramurthy, S., Subramaniam, S., and Ramasamy, C., 2015. Distilling big data: Refining quality information in the era of yottabytes. *The Scientific World Journal*, 2015.

Razzaque, M. A., Bleakley, C., and Dobson, S., 2013. Compression in wireless sensor networks: A survey and comparative evaluation. *ACM Transactions on Sensor Networks (TOSN)*, 10(1):1-44.

Reed, D., Sprony, M., Longley, D., Allen, C., Grant, R., and Sabadello, M., 2018. Decentralized identifiers (DIDs) v0. 11 data model and syntaxes for decentralized identifiers (DIDs). W3C. *W3C, Cambridge, MA, USA, Tech. Rep*.

Ren, Y., Zeng, Z., Wang, T., Zhang, S., and Zhi, G., 2020. A trust-based minimum cost and quality aware data collection scheme in P2P network. *Peer-to-Peer Networking and Applications*, pages 1-24.

Reps, T., Horwitz, S., Sagiv, M., and Rosay, G., 1994. Speeding up slicing. *ACM SIGSOFT Software Engineering Notes*, 19(5):11-20.

Rodríguez, C. C. G. and Servigne, S., 2013. Managing Sensor Data Uncertainty: a data quality approach. *International Journal of Agricultural and Environmental Information Systems (IJAEIS)*, 4(1):35-54.

Rosenthal, S., Mohammad, S. M., Nakov, P., Ritter, A., Kiritchenko, S., and Stoyanov, V., 2019. Semeval-2015 task 10: Sentiment analysis in twitter. *arXiv preprint arXiv:1912.02387*.

Rubin, V. and Lukoianova, T., 2013. Veracity roadmap: Is big data objective, truthful and credible? *Advances in Classification Research Online*, 24(1):4.

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

Samuel, H. and Zaiane, O., 2018. MedFact: Towards improving veracity of medical information in social media using applied machine learning. In *Canadian Conference on Artificial Intelligence*, pages 108-120. Springer.

Sanyal, S. and Zhang, P., 2018. Improving quality of data: IoT data aggregation using device to device communications. *IEEE Access*, 6:67830-67840.

Shahnawaz and Astya, P., 2017. Sentiment analysis: Approaches and open issues. In *2017 International Conference on Computing, Communication and Automation (ICCCA)*, pages 154-158.

Shannon, C. E., 2001. A mathematical theory of communication. *ACM SIGMOBILE mobile computing and communications review*, 5(1):3-55.

Singh, J. P., Rana, N. P., and Dwivedi, Y. K., 2019. Rumour Veracity Estimation with Deep Learning for Twitter. In *International Working Conference on Transfer and Diffusion of IT*, pages 351-363. Springer.

Srivastava, P. and Mostafavi, A., 2018. Challenges and opportunities of crowdsourcing and participatory planning in developing infrastructure systems of smart cities. *Infrastructures*, 3(4):51.

Tarmizi, F. A. A., Tan, P. X., Sharif, K. Y., and Kamioka, E., 2019. Online news veracity assessment using emotional weight. In *Proceedings of the 2019 2nd International Conference on Information Science and Systems*, pages 60-64.

Tekiner, F. and Keane, J. A., 2013. Big data framework. In *2013 IEEE International Conference on Systems, Man, and Cybernetics*, pages 1494-1499. IEEE.

Turner, S., 2004. Defining and measuring traffic data quality: White paper on recommended approaches. *Transportation research record*, 1870(1):62-69.

Tzeng, G.-H. and Huang, J.-J., 2011. *Multiple attribute decision making: methods and applications*. CRC press. Vosoughi, S., Mohsenvand, M. and Roy, D., 2017. Rumor gauge: Predicting the veracity of rumors on Twitter. *ACM transactions on knowledge discovery from data (TKDD)*, 11(4):1-36.

Wibowo, S. and Sandikapura, T., 2019. Improving Data Security, Interoperability, and Veracity using Blockchain for One Data Governance, Case Study of Local Tax Big Data. In *2019 International Conference on ICT for Smart Society (ICISS)*, volume 7, pages 1-6. IEEE.

Wibowo, S. and Sumari, A. D. W., 2020. The Utilization of Blockchain for Enhancing Big Data Security and Veracity. In *Combating Security Challenges in the Age of Big Data*, pages 157-187. Springer.

Wu, K., Yang, S., and Zhu, K. Q., 2015. False rumors detection on sina weibo by propagation structures. In *2015 IEEE 31st international conference on data engineering*, pages 651-662. IEEE.

Yin, S. and Kaynak, O., 2015. Big data for modern industry: challenges and trends [point of view]. *Proceedings of the IEEE*, 103(2):143-146.

Zaparniuk, J., Yuille, J. C., and Taylor, S., 1995. Assessing the credibility of true and false statements. *International Journal of Law and Psychiatry*.

Zhang, J., Wong, J.-S., Pan, Y., and Li, T., 2014. A parallel matrix-based method for computing approximations in incomplete information systems. *IEEE Transactions on Knowledge and Data Engineering*, 27(2):326-339.

Zhou, Y., De, S., Wang, W., and Moessner, K., 2016. Search techniques for the web of things: A taxonomy and survey. *Sensors*, 16(5):600.