Exploring and Evaluating the Impact of the Veracity of Big Data Sources

Bayan Hazaa AlDoaies
Faculty of Computing and Information Technology
King Abdulaziz University
Jeddah, Saudi Arabia

Anhar Maatoq Ashi
Faculty of Computing and Information Technology
King Abdulaziz University
Jeddah, Saudi Arabia

Fahd S. Alotaibi
Faculty of Computing and Information Technology
King Abdulaziz University
Jeddah, Saudi Arabia
Email: fsalotaibi [AT] kau.edu.sa

***Abstract*— Exploring the data veracity recognized as one of the significant challenges in the big data sources. The veracity is an aspect of big data that deals with correctness and trustworthiness. In fact, it's hard to obtain the truth of big data in Website, and social media platform which has a rich volume of information with surpassing conflict. In this paper, we proposed a comparison between veracity models which have studied in both Website and Twitter platform. In a comparison, the several criteria had been demonstrated, as verification aspects, working flow, trust score, accuracy, security, usability, and time-consuming. Thus, each model has its attributes and an approach to extract the data veracity based on the used platform. Consequently, the models' approach defines the way of evaluating the impact of the credibility on the big data.**

***Keywords*— Big data, Veracity, Data trustworthiness, Credibility, Trust-score, Accuracy, Verification aspects.**

I. INTRODUCTION

"Learning to trust is one of life's most difficult tasks."
– Isaac Watts.

With the advance of Information Technology, the World Wide Web plays a key role to obtain, gather and process the data by providing different sources. Nowadays, the Website and Twitter platform consider as the main sources of information that people look for it. These platforms allow the user to extract and share the data over a broad geographic region without the temporal and spatial limitations. Also, it simplifies the release of information on the large scale. Furthermore, the vast amount of data being generated with surpassing the processing and analytical ability within the time needed is called big data. The large datasets become the difficult challenge not only to generate enormous volume but also to store, search, share, visualize, analyze and test. Moreover, in dealing with the big amount of data, it will be tough to meet the most 5Vs requirements to validate it and to ensure no issues persist on data quality or performance which are Variant, Velocity, Volume, Value, and Veracity. Further, there are many researchers interest to work on Velocity, Volume, Value and Variety of information over the web, while the most critical dimensions have missed the Veracity of extracting information. Consequently, Veracity means preservation the main attributes of raw data and it necessary to trust the information to use it in various domains for several goals. Also, it refers to the quality of trustworthiness data or the correctness of the data. By the same token, the trustworthiness, credibility, truth, truthfulness, integrity, and correctness all are synonymous for the data veracity. Further, it's difficult for an individual to take the conscious decision about the veracity level of the readable data. Wherefore, there are many models and techniques to extract a high-quality data. According to our main contribution, this paper will address the lack of comparative studies of veracity models by collecting different techniques of the Website and Twitter platform and study it under different criteria as verification aspects, security, accuracy, and more.

This paper organized as follows. First, we review literature review in Section II. Section III discusses the veracity challenge, its models, and the comparison criteria. In Section IV the comparative study takes place. Significant Discussion presents in Section VII.

II. LITERATURE REVIEW

The related work to our approach falls to explore the impact of the veracity of big data sources such as Websites and Twitter by collecting different techniques aims to extract the trustworthiness data.

Several mechanisms and techniques have been proposed to measure the veracity of big data, some of the techniques depend on information provenance as Veracity Ontology model which recommended by [8]. This model embedded in web agent in a secure way by trusting information provenance. Moreover, it uses Semantic Web technologies which defined as a proposition at information level to validate information content. This model may lead to mistake if used with a personal and social platform. Further, it needs expert support and brings heavy workload. Also, [15] aimed to trust a data on the web. Hence, the researcher designed models depend on provenance information and information consumer opinion. These models are Trusted model, Trust Automatic Assessment, and methods to utilize trust assessments. For precise analysis and effective decisions, [9]; [6] and [11] used trust score to assure information veracity and provide the user with data need in credibility status from an accurate Website. First, [9]; developed Data Provenance Trust model. This model based on different factors which influence the veracity and based on

these criteria assigns trust score to both data and data providers. Furthermore, they tested their approach and reached to the efficiency and feasibility of their model and the high ability to work on a similar data and with unexpected mistakes. As well, [6] they proposed Assuring Approach base on two key elements which are trust score and the confidence policy. The suggest approach not limited to particular type of application or domain, but it's general and can use it in different fields. Quality Knowledge-Based Trust (KBT) is a model based on the probability that it contains the correct value of a fact which designed by [11]. The researchers proposed this new metric to estimate web source trustworthiness. They used many techniques to achieve this model as information extraction and inference in a probabilistic model. From the most popular techniques used with data veracity is Truth Finder. [16] and [19] proposed this technique in their studies. However, they used the interdependency amongst Websites and their information to identify the correct fact in the trustworthy site. Moreover, developers reached to that Truth Finder obtains high accuracy for finding fact and veracity information in the trusted Website. Also, it identifies the web source that has more precision data. In the [4] paper, the main challenge is estimating accuracy. The researcher proposed the Fact-Finder approach which gives a result of validation approach by reduce expected errors and gives the best quality performance on average for estimating the veracity of data with small ground trust. Moreover, the number of data on the web has been growing up, and a lot of information are not trusted people can rely on it. So, [20] in their paper proposed to reduce a distance between the correct data and the overall description through estimating the accuracy and cover data. The efficiency of the real data set clarifies reflected on their experiment. [18] focused on increase the veracity of web content to avoid extract untruth and the conflict information. So, they developed an algorithm based on Link Density and Statistic which embedded in a search engine. The algorithm focused on semi-structured text for all domains. The researcher reached that a process reduces the complexity, transmission quantity, and need a lot of effort before start working on search engine interface, and it's suitable for expert user and the worker. Furthermore, to confirm truthful information [10] proposed eight linked data quality metrics and techniques as Revisor Sampling and Bloom Filters, and reached to how quality metrics and techniques improve user's viewpoint for data. Additionally, estimating data veracity in social networks and the Website was the key aspect discussed by [3]. They proposed the VERA which is an approach of a Web-based and Twitter platform that supports the data extraction from the Web textual and micro-text from Twitter to estimate the veracity. It contains four layers which help the user to explore the truth, understanding how data estimation, and how the system return score. So, with this layers in VERA, they can assure the data veracity.

In the recent years, the social media becomes the popular platform to share data over a broad geographic region and to demonstrate how the information that spread through the social media can be trusted or not. There is multiple researchers studied data veracity on the most public platform which is Twitter. [21]; [14]; and [2] focused on their studies on the social network especially Twitter platform. However, when estimating the credibility, most of prior worked focused on general tweets as independent of each other. [21] studied how to trust users and tweets on Twitter and how to treat Twitter as trust source of information and news. To do that, they proposed Novel Topic Focused Trust model and Trust Propagation for estimating a trustworthiness scores. As well, [14] aimed to develop an Autonomous Message Classifier to filter data on Twitter. They determined over 80 standard measurements and designed GUI to determine the trustworthiness score and identified an Automatic Measurement for any communication. They reached to that a model is more accurate than other prior worked, but it's still manual, time consumer, and need to rebuild to be automatic. As well, [7] evaluated the veracity of social media networks from applications ranging to predict product demand. Each classifier is trained on each collection of input data transform it and test it. They define the user as being accurate if and only if he\she sends a message assumption an event is real or fake, and that claim matches the actual outcome of an event. The truth determined based on time, distance and trust metrics. Moreover, [2] proposed Quantitative Mechanisms indicators as (i) topic diffusion, (ii) geographic dispersion, (iii) spam index. These mechanisms based on tweets themselves to determine a level of accuracy and veracity of published Twitter topics. They reached to that quantitative indicators are particularly useful measurements to appreciate and compare the veracity level for most topics, and to estimate election campaign data. [17] assessed the veracity manually of the open source information such as Web, Twitter and all information available in the network. They assess by interviews and questioners use a ranking scale that used in assessment needs. Hence, with this assessment they can conclude that automation and systematization of the veracity assessment would be highly beneficial. [1] used the Crowdsourcing Mechanism which has a technique of tag me application to extract tweet from Twitter and display it, depending on sentiment analysis. As well, the collected data has been evaluated with the verified data set to prove the accuracy.

According to our review of the cited literature, we indicate that a source information and content features are helpful and fundamental points to distinguish content veracity in both Website and Twitter platform, and avoids the only opinion in the evaluation. Accordingly, for choosing mechanisms to evaluate data credibility, there are many criteria must be considered which will define in Section IV.

III. VERACITY EXPLORING

The concentration of this paper is exploring the veracity of big data in Twitter platform and Websites. As we saw in the literature review, there are many models proposed to verify the data veracity on both resources, and each model has a particular way to extract a trustworthiness data which effect on enterprise's decision. The significance of data integrity in all life fields and diffusion of veracity models make it worthy to have structure comparison paper to study the difference between veracity models. However, this paper will make a

contribution, and facilitate to have a complete view of each model and assist in answering our research questions which are about, what the difference between these mechanisms and how each one effect on the data veracity on both web sources and Twitter platforms.

In this section, we will briefly describe the veracity challenges then will describe each model approach to compare them under different criteria as the verification aspects, working level, trust score, accuracy, security, usability and time-consuming.

A. *Veracity challenge*

One of the major barriers is that data be unreliable if it is tendentious, misguiding, mistaken, careless or antiquated. Moreover, with the propagation of the conflict data across different sources the evaluation of data veracity becomes fundamental search point [5]. However, based on [9] there is four aspects effect on veracity models: (2) data similarity, (2) path similarity, (3) data conflict, (4) data deduction. In the following, there is a brief description of each factor.

1. *Data Similarity*

Data similarity refers to the semblance of data features like location, date, size, and more. However, the huge similarity of data will not guarantee the data trustworthiness.

2. *Path Similarity*

The path similarity impacts must account it when computation data similarity and veracity models.

3. *Data Conflict*

Data conflict indicates the incompatible characterization or facts about the same structure or event. Besides, it has an adverse effect on the trustworthiness of data.

4. *Data Deduction*

The trustworthiness data is significant to guarantee high-quality decisions.

B. *Various types of veracity model in Website resources*

In the following, it's offering different models which bring out the veracity of big data.

1. *The Veracity Ontology Model (VO)*

The Veracity Ontology model works at information level to verify the veracity of information provenance and content by using Semantic Web technologies. The trustworthiness based on social and rational variables, which both complementary to each other. The VO depends on the concept of the proposition, agent, and trustworthiness. So, the proposition must identify to estimate the trustworthiness in the piece of information and define accurate agent who will evaluate the proposition. To ensure about proposition veracity, the external piece of information must exist to support or falsify the veracity rate. Moreover, the VO model securely distributed by using a digital signature to ensure that web provenance can't easy fabricate it or makes an edit on web resources. Also, it has a high level of agent credibility, proof assertion, reliable and secure assert, all these aspects not available on other models. Besides, it needs proficient assessment and brings heavy workload.

2. *Trust Assessment Model*

Trust Assessment model depends on provenance information and opinion of the information consumer to represent trustworthiness of the data on statement level. This model is automatic and manages trust value in an efficient manner by using trust function. In contrast to other models, it has a uniform approach for assisting data trustworthiness and controls access to the assessment. Also, it focuses on verifying a veracity of data released on web source, instead of publishers.

3. *Data Provenance Trust Model*

Data Provenance Trust model bases on different factors which influence the veracity and based on these criteria assign trust score to both data and data providers. The trust score plays as the essential key to assign the judging rate to the data veracity based on what and why the data used. It requires high assurance data integrity to extract good quality data. Moreover, it works well with both unintentional errors and malicious attacks without collusion. Besides, the approach work efficiency with large dataset size, it takes less than one second to compute trust score, and high efficiency appears when dealing with high trust score which aims to pass untrusted data on a system.

4. *Assuring approach*

The Assuring Approach focuses on data provenance to determine trust score that considers as a key value to trust data. However, trust score used for data comparison or ranking with interdependency between the data provider and data item with the assessment of trust score. Provenance information is essential for ensuring and improving data trustworthiness. Also, it provides confidence policy to determine which data use it for the specific task and dynamically intercepts access to the inquiry result depend on trust score, and it may be difficult. Moreover, very significant point focused on this approach is the security of data provenance because it considers as the key point for determining data trustworthiness. As a consequence, a Digital Signature and Cryptography techniques used to deny such virulent attacks. Besides, XML language uses for encoding and securing provenance information. This approach general and can use it in different fields.

5. *Knowledge-Based Trust (KBT)*

Knowledge-Based Trust works on source level and relies on exogenous signals to define an accuracy of web source by measure a probability for the correct value of a fact. It differentiates between two types of error which are incorrect fact and incorrect extraction. Hence, the approach provides accurate respect of the source reliability, correctness of

extraction, and the quality of extractor. Moreover, for improving web source quality, the KBT provides the valuable additional signal.

6. Truth-Finder

The Truth Finder extracts data from Websites which have huge numbers of data that people rely on it. This model used to reduce the distance between the true data and the observed data by determining the accuracy and the coverage to data to find the trustable scores.

7. Fact-Checking Model

The Fact-Checking model used to improve the quality and the performance of method by using the ground truth. In detail, the Fact-Checking used to estimate the true data. The technique used to compute the accuracy of the selected methods and evaluate it when biased, also when small ground of truth exists. However, it considers costly and time-consuming. As an illustration, the accuracy computed in Truth Finder method with 11 claims and 100% ground of truth data in the result is 11/11.

8. Link Density and Statistic Algorithm

The algorithm focused on semi-structured text on the web page to increase the veracity of web content and avoid extract untruth and the conflict information. Moreover, it embedded in a search engine which offers easy access to web pages to make information extraction an ordinary mission. Besides, this model is global and applicable for the most Website, and it's very suitable for using by expert or other users. Moreover, it reduces the magnitude of data transition and complexity.

9. Linked 'Big' Data

The Linked Big Data establishes eight linked data quality metrics and techniques as Revisor Sampling and Bloom Filters to achieve a high veracity of the data. The Revisor Sampling is a statistics-based technique of randomized sampling which finds the solution for time consuming. In contrast, Bloom Filters used to map and compare element in a set. The quality metrics improve a consumer viewpoint and improve dataset's value. Moreover, the two techniques Revisor Sampling and Bloom Filters can use with big dataset whose time complexity become ungainly.

10. VERA Model

VERA is an approach used to estimate the veracity extracted data from Website and Twitter. In another word, VERA is an architecture that described in four layers: information extraction, data fusion, truth discovery and the visualization, and explanation. As an illustration, in the first layer, the information extraction depends on the type of the data resource. For an instant, TextRunner used to extracts data from the Website, DeepDive extracts data from the document, and TweetLE extracts it from Twitter. Furthermore, data fusion layer which transforms extracted data into claims. More importantly, the truth discovers layer which is used to score the veracity depends on the claims if it's true or false. Finally, the visualization and explanation layer contains GUI which helps to explore the truth, understand the data, and return scores.

C. Various types of veracity model in Twitter platform

In the following, there is offering different models which bring out the veracity of big data in Twitter.

1. Topic Focused Trust Model

The Topic Focused Trust model elaborates on how to trust users and tweets on Twitter and how to treat Twitter as trust source of information and news. 4e: first, most models focus on estimating the truthfulness on general topic while a Topic Focused Trust model concentrates on user's interest. Second, use trustworthiness news' reports to indicate the trustworthiness of tweets present contextual symmetry in textual, locative and temporal features. The third feature uses semantic and contextual information with social platform data for trustworthiness diffusion. It's automatically used a Novel-Trust Evaluation Mechanisms for rating the similarity of topic focused tweets. On the other hand, the Novel Iterative Trust Propagation Algorithm define the relationship between the contextual and social of tweets to estimate the trustworthiness. So Far, a malicious user attack only considers in this model, with future consideration plan for including random, opportunistic, and insidious attack behaviors. Moreover, the model works stably with different languages across countries, and this reflects the effectiveness and robustness of this model.

2. Autonomous Message Classifier

Autonomous Message Classifier to filter data on Twitter with over 80 standard measurements proposed and designed GUI. Moreover, the machine learning classifier is run over these measurements to determine the trustworthiness score at a variant time and identify an automatic measurement for any communication. These measurements are Decision Tree, Random Forest, and Logistic regression. The accuracy rate of the models reached to 96.6%, according to two reasons, (1) introduce a user's present and past behavior, (2) the process of gathering tweet/user features to rumor properties. The classification of the tweets into rumors is a manual and time-consuming task and requires significant effort and much time. In contrast, no information mentioned about the confidence level of the outcome.

3. *Data transformation and User Classification Trust Model*

The user in social media assumed to be as a classifier that takes various information classifies an event as it real or hoax. It referred to the trustworthiness of user by classification the event of his/her to the actual event. In determining the trustworthiness by the method, there are two steps, first, the data of user profile transformed into the feature vector. Then, classifiers applied vector feature to determine whatever post are trusted. Finally, filtering the system to determine which message are more trustworthy. Further, to claim the user if

accurate or not if and only if he/she sends the message claiming an event is real or fake and that matches real, or hoax come event the system that developed is 75% accurate. Besides, resulting in some users being misclassified as trustworthy and noise still getting through the filter.

4. Quantitative Mechanisms

The Quantitative Mechanisms indicators as topic diffusion, geographic dispersion, and spam index, based on tweets themselves to determine a level of accuracy and veracity of published Twitter topics. The proposed mechanisms compare information extracted from tweets with information from formal data sources, and this way considers a best to determine the veracity of tweets. However, this process may be time-consuming. Besides, quantitative indicators are particularly useful measurements to appreciate, evaluate, and compare the veracity level for most topics and estimate election campaign data.

5. Open Source Information (OSINF)

The OSINF is the framework which helps to assess the integrity of a large amount of data in short time. OSINF has a very distinct quality and making a challenge in evaluating the trustworthiness demand. Furthermore, in the method, it assumes that there is a significant number of OSINF component either assessed or not. For the purpose of the primary challenge, the method is automation to obtain an assessment with high quality, and an accuracy in short time instead of spending time working with irreverent data.

6. Crowdsourcing Model

The crowdsourcing model used as a solution to the problem of verifying the veracity of data in Twitter. The main idea of the solution found by using sentiment analysis in every piece of text. Obviously, it's an app tagged tweets of the user as per the sentiment, and the tagged tweets compared to verified data set. In additionally, the measure of accuracy in this model done by ROC and Bayesian curve.

IV. COMPARATIVE CRITERIA

Depending on the collected data about models, the different criteria demonstrate to measure the efficiency of veracity model. In the following, there is a brief description of the comparison criteria.

1. Verification aspects

Verification aspects define which part of the model will work on it to verify data veracity.

2. Working Level

The structure level which a model is working on it, for instance, statement, source, rumor, or information.

3. Trust Score

The key information for ensuring data trustworthiness based on which data the user may use and for what purpose [9].

4. Accuracy

The probability of correct value or valuable data that contains in web source [11], and in this comparison, it measures the models' performance.

5. Security

Manage authentication internally and externally, client to node encryption, and transparent data encryption and data auditing [13].

6. Usability and Time-consuming

This criterion measures the simplicity of the model to fit a user need. Also, it studies the time taken to ensure the data veracity.

V. COMPARTIVE STUDY

In the previous section, the different models have been discussed, and therefore to achieve the paper goal the comparative study which considered as the result of this paper will take place. The comparison categorized based on various criteria which mentioned in Section (IV).

A. *Verification aspects of data veracity*

1. *Website models*

As we see in Table I, each model based on various aspects to verify data veracity. Veracity Ontology model based on the proposition (or a semantic statement) which extracts from web content to proof it by the trusted agent. In contrast, Trust Assessment model depends on Data provenance and aggregation the trust rating of consumer judgment. Moreover, Data Provenance Trust model assures the veracity of data provenance by assigning trust score for both data item and data provider. Like all previous models, the Assuring Approach model depends on data provenance to gives evidence about how and where the data is generated and based on confidence policy. KBT model evaluates data veracity by measure the facts' probability on data provenance. In contrast, Link Density and Statistic Algorithm based on web content to extract the truth data. In the same way, the input or parameters which extracted from Fact-Checking model and Truth Finder model depend on different aspects, as the number of provenances and the data content which used to determine the quality performance. As a matter of fact, in verifying the veracity of extracted data in VERA model using various information extractor depending on the set of resources. The extractors are, (1) TextRunner is an open information extraction system which extracts information from a Website, (2) the deepdive predefined extractor takes as input from collection textual document and find a relationship between them. Furthermore,

Linked 'Big' Data based properties of the dataset and data resources.

TABLE I
VERIFICATION ASPECTS OF DATA VERACITY ON THE WEBSITE

Models	Aspects					
	Proposition	Data provenance	Consumer opinion	Confidence policy	Data content	Data provider
VO Model	×		×			
Trust Assessment Model		×	×			
Data Provenance trust Model		×			×	×
Assuring Approach		×		×		×
KBT		×				
Link density and statistic					×	
Fact-Checking Model		×			×	
Truth Finder		×		×	×	
VERA Model		×				×
Linked 'Big' Data		×			×	

TABLE II

VERIFICATION ASPECTS OF DATA VERACITY ON THE TWITTER (A)

Models	Aspects			
	Textual similarity	Spatial similarity	Temporal similarity	Diffusion
Topic-focused trust Model	×	×	×	
Quantitative Measure				×
Autonomous Message Classifier				
User trusts Modeling		×	×	
OSFIN Model			×	
Crowdsourcing Model				
Vera Model				

TABLE III

VERIFICATION ASPECTS OF DATA VERACITY ON THE TWITTER (B)

Models	Aspects			
	Spam	Geographic spread	linguistic	User's behavior
Topic-focused trust Model				
Quantitative Measure	×	×		
Autonomous Message Classifier			×	×
User trusts Modeling		×		×
OSFIN Model				
Crowdsourcing Model				×
Vera Model				×

2. Twitter models

The aspects listed in Table II and Table III belong to Twitter model for trusting tweets/users. Consequently, Topic-Focused Trust model rates the trustworthiness tweets/users focused on the user interest topics by evaluating heterogeneous factors (I) textual similarity, (iii) Spatial Similarity, (iii) temporal similarity. Accordingly, the User Trusts modeling define the trustworthiness of user by collecting the data from a user profile and compare the user reaction to an event to the actual value of the event. Furthermore, demonstrate how information about an event in real-world impact when spread through a social media networks based on diverse of geographic regions, threat scenarios and time investigated. Moreover, the veracity of Twitter topics bases on associated tweets that in the Quantitative Measure. This model uses three different measurements, topic diffusion which computes the fast of spread information through Twitter, geographic dispersion measures the geographic expansion, and spam index calculates the effect of repeated tweets by the same user. Autonomous Message Classifier based on two aspects, linguistic of topics, user's present and past behavior. Equally important, the OSFIN model mentioned the data extraction in model by following three steps, first, the need to assert whether a source is the originator of the information or only rely on it. Second, check the history associations of a source. Third, follow it over time. Furthermore, The Crowdsourcing model uses the TAG ME! app to take data from Twitter states or each time user tweet or tags the tweets and mapping every tweet to the user by the ID. Also, in VERA model one of extractor TwitIE applied to set of tweets collected from users and find relation extraction between tweets, real-time not supported.

In summary, Table I approves that data provenance is the most important aspects of evaluating data veracity and 8 from 10 models based on it. In contrast, proposition, consumer opinion, and confidence policy consider as lowest level aspects that model used in evaluation, within reason of difficulties as the proposition or not assure it veracity as user judgment. In contrast of Website, the most public evaluation aspects of Twitter models that shown in Table II and Table III is user's behaviour, because the Twitter platform is based completely on user viewpoint and behaviour. Nevertheless, it intractable to depend completely on consumer's opinion and ignore the other aspects even with verification features which placed in the Twitter platform. This feature doesn't guarantee the trustworthiness of users/tweets because everyone can verify and increase their followers' account by paying some money.

B. *Working Level*

There are four models clarified its working level as VO Model that estimates the proposition veracity at information level, while Trust Assessment model defines the trustworthiness of data on statement level. KBT model is working on source level to guarantee the data veracity. Finally, Autonomous Message Classifier aggregates all feature on rumour level.

C. *Trust Score*

There are multiple models use trust score as a key notion to represent the trustworthiness level of mentioned aspects in Table I, Table II, and Table III each model has its approach to generate the trust score, as will see in the following. Trust score in Data Provenance Trust model assigns to both data item and data provider depends on the amount of valid data it has supplied. Besides, data users can take the decision even use the rated information, or to further verify information. Trust score considered four factors that influence trustworthiness. Moreover, assuring approach used the trust score for data comparison or ranking, and it ranges from 0 as the lowest level of trustworthiness and 1 conversely. However, Trust score based on data provenance and can be acquired by using multiple factors as trustfulness of data provider and a way of data aggregated with an interdependency property between them. Moreover, trust score will assign for confidence policy to restrict access to the inquiry result dynamically and it ranges between $[q_{max}, q_{min}]$. In additionally, it based on which the data to be use and for what purpose. Also, trust function uses in Trust Assessment model to determine truth value of data provenance and the information of consumer opinion. In contrast, Topic-Focused Trust model allocates trustworthiness score for tweets/users to estimate it credibility. The score of tweets interdependency with the reality of things happen, otherwise for users determine it by the user's tweets. Furthermore, VERA model presents the complete list of sources which support the corresponding data, their trustworthiness score computed by truth discovery layer. Moreover, the third layer in VERA approach that responsible for executing various truth discovery methods and determine which data is true or false by computing the veracity score and estimating the trustworthiness score of the sources. Finally, Autonomous Message Classifier assigns different critical attributes to determine the trustworthiness score at the variant time.

D. *Security*

Security is a very significant issue that provides a secure environment for ensuring trustworthiness. There are many models take this issue in its account as VO model, Data Provenance Trust model, Assuring Approach, and Topic-Focused Trust model. Digital signature (DS) used by VO model and Assuring Approach model by associating a DS unique key to each assertion trust and match it with a particular agent. It is used to ensure the data not falsified and constructs the steady rule for assuring process. Another technique used on Assuring Approach model is Cryptography that prevents the malicious attack and used XML for encoding source information. Moreover, Trust scores in Data Provenance trust model not only for assure the veracity of data provenanc. Nevertheless, it helps to deal with both unintentional errors and virulent attacks. Also, it ignores the high value of trust score and assigns the data items with "Newly arrive" value until another source provides report a similar data. Finally, Topic-focused trust model concentrates on the malicious attack without disguise whenever it has the chance.
The security techniques aren't confined to the mentioned ones, but there are different algorithms based on various properties which not used it on all models. The security techniques as Full disk encryption (FDE), RBAC Authorization approach, message authentication code (MACs) [12], and Firewall techniques.

E. *Accuracy*

There are various models compute its accuracy percentage and its range from 96.6% to 75%. Thus, as shown in Figure 1 each algorithm of Autonomous Message Classifier model has different accuracy rate, the best performing is for Decision Tree algorithm with 96.6%. Random Forest algorithm follows with almost 90% accuracy rate. The lowest accuracy rate achieved of the three models is for Logistic Regression with accuracy close to 82.8%. Moreover, Link Density and Statistic algorithm meet the requirements of data extraction with nearly 95% accuracy rate. Furthermore, the Crowdsourcing model measures the accuracy confidence with 88.85% by using different classification methods which are ROC curve, Bayesian predictor function, MLE, and MAP. By the same token, ROC curve used to measure the performance and accuracy rate of User Trusts modeling and it nearly 75%.

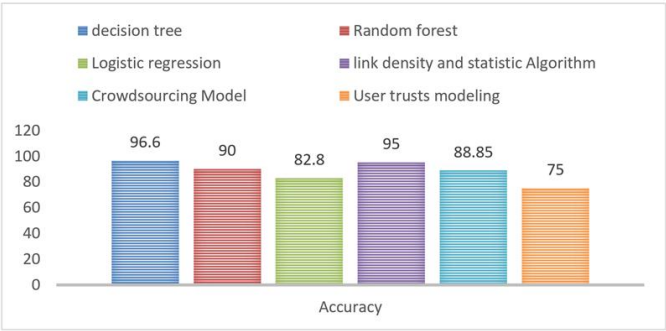

Figure 1: Accuracy rate for veracity models.

F. *Usability and time-consuming*

The model to be efficient it must meet consumer needs as save the user time and be useful for everyone not only for professional workers. Some models address these issues, in the following, there is brief comparative in how the models deal with the consumer needs. VO model needs proficient assessment and brings heavy workload and not useful for end user. In contrast, Link Density and Statistic Algorithm is compatible with most new Websites and very useful for expert or end user. Moreover, it reduces the magnitude of data transition and complexity. Moreover, Autonomous Message Classifier model requires much effort and time to classify the tweets into rumors because it is a manual and time-consuming task. Consequently, these models need to be automated to classify the tweets dynamically. Besides, VERA model results visualization and explanation consist of a set of Web user interface to support the system usability by easy exploring the results of the truth discovery and understanding how to compute the estimation of the veracity by the system. Another significant result is Explanation which accomplished in VERA through APIs whereas result visualization renders the output of the truth discovery process to ease user exploration and interaction with the system. Thus, in Crowdsourcing model, Tag Me is an application consists the home page or the login screen where the user can interact with, and the result with the score displayed. Also, the dynamic update of the leaderboard will not consume the time. Also, OSINF model can assess the veracity and analyze the large amounts of data within a short amount of time. By the same token, Trust User model takes information either by pinpointing the event's location or time and supports the real-time event detection is possible to prevent time-consuming. Otherwise, Linked 'Big' Data model solves the problem of time-consuming which appear in the last model by providing the Revisor Sampling technique. Furthermore, the two techniques Revisor Sampling and Bloom Filters can be used with big dataset which its time complexity become intractable. Also, Data Provenance Trust model works efficiently with large dataset size. As a consequence, it takes less than one second to compute trust score. KBT is a dynamic algorithm to estimate the granularity level for web source. Finally, Quantitative Measure prevents time-consuming by focus on tweets associated with the topic rather than the source.

In summary, for the end user, the most suitable models which are Linked Density Statistic Algorithm, VERA model, Crowdsourcing model, and KBT either it has GUI that allow the consumer to interact with or it's work dynamically. In contrast, VO model, and Autonomous Message Classifier model are suitable more for the expert user. Otherwise, most of the studied models prevent time-consuming by reducing data complexity and improve the dynamic used while Autonomous Message Classifier model still work manually which waste users' time.

VI. SIGNIFICANT DISCUSSION

Commenting on the previous comparative study, track images and multimedia which embed in site/tweets content didn't take into account all studied models in this paper. This approach is very significant by analyzing images and multimedia properties and checks if it has been posting before that or fabricates it in some way. Moreover, embedded model in search engine considered as the most useful way for everyone to use. For that, we think to merge between the studied models and the plugin program which will be embed in the web browser to notify the user. A notification will be done via the pop-up message that appears when the user searches on the Website or any data source that contain untruth information.

VII. CONCLUSION AND FUTURE WORK

As we have seen in the paper, there are several models have been evaluated for exploring the veracity of big data in Website and Twitter. In this paper, we considered the comparison between several models using different criteria. As an illustration of criteria, verification aspects of Website models are data provenance which was the most used, also proposition, consumer opinions, confidence policy and data provider had been chosen with the different rating. In contraction for Twitter models, the aspects of data verified were by user behavior which the most used and the data selected by the diversity of time and location. In addition, the working level of models can be a source, information, statement, or rumor level. Moreover, a generating way of trust score considers as comparative criteria to make the data comparison or ranking the trustworthiness level. In the security criteria, it is important to provide a secure environment to explore data veracity by using different security techniques. In the same token, the accuracy performance of the model is ranging from 96.6% to 75%. For usability, there are varying levels of models' complexity. In the same way, the models have different control methods to manage a time and prevent time consumed.

This comparative study guides us to do a deep research to verify veracity in various platforms as social multimedia. In the future, we plan to evolve a multimedia approach which assumes the veracity level of video and images in YouTube, Snap Chat, and Instagram platforms by analyzing media content and determine the credibility level. Also, as mentioned in Section (VI), a proposed notification's plugin program will consider during the future work.

REFERENCES

[1] Bhoomika Agarwal, Abhiram Ravikumar, and Snehanshu Saha. A novel approach to big data veracity using crowdsourcing techniques and bayesian predictors. 2016.

[2] Kumar TK Ashwin, Prashanth Kammarpally, and KM George. Veracity of information in twitter data: A case study. In *2016 International Conference on Big Data and Smart Computing (BigComp)*, pages 129–136. IEEE, 2016.

[3] Mouhamadou Lamine Ba, Laure Berti-Equille, Kushal Shah, and Hossam M Hammady. Vera: A platform for veracity estimation over web data. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 159–162. International World Wide Web Conferences Steering Committee, 2016.

[4] Laure Berti-Equille. Data veracity estimation with ensembling truth discovery methods. In *Big Data (Big Data), 2015 IEEE International Conference on*, pages 2628–2636. IEEE, 2015.

[5] Laure Berti-Equille, Mouhamadou Lamine Ba, and Hossam M Hammady. Veracity of big data: Challenges. 2015.

[6] Elisa Bertino and Hyo-Sang Lim. Assuring data trustworthiness-concepts and research challenges. In *Workshop on Secure Data Management*, pages 1–12. Springer, 2010.

[7] Todd Bodnar, Conrad Tucker, Kenneth Hopkinson, and Sven G Bilén. Increasing the veracity of event detection on social media networks through user trust modeling. In *Big Data (Big Data), 2014 IEEE International Conference on*, pages 636–643. IEEE, 2014.

[8] Grégoire Burel, Amparo E Cano, Matthew Rowe, and Alfonso Sosa. Representing, proving and sharing trustworthiness of web resources using veracity. In *International Conference on Knowledge Engineering and Knowledge Management*, pages 421–430. Springer, 2010.

[9] Chenyun Dai, Dan Lin, Elisa Bertino, and Murat Kantarcioglu. An approach to evaluate data trustworthiness based on data provenance. In *Workshop on Secure Data Management*, pages 82–98. Springer, 2008.

[10] Jeremy Debattista, Christoph Lange, Simon Scerri, et al. 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, 2015.

[11] Xin Luna Dong, Evgeniy Gabrilovich, Kevin Murphy, Van Dang, Wilko Horn, Camillo Lugaresi, Shaohua Sun, and Wei Zhang. Knowledge-based trust: Estimating the trustworthiness of web sources. *Proceedings of the VLDB Endowment*, 8(9):938–949, 2015.

[12] Ajit Gaddam. Securing your big data environment. In *Black Hat USA 2015*. Black Hat, 2015.

[13] Naveen Garg, Sanjay Singla, and Surender Jangra. Challenges and techniques for testing of big data. *Procedia Computer Science*, 85:940–948, 2016.

[14] Georgios Giasemidis, Colin Singleton, Ioannis Agrafiotis, Jason RC Nurse, Alan Pilgrim, and Chris Willis. Determining the veracity of rumours on twitter. 2016.

[15] Olaf Hartig. Trustworthiness of data on the web. In *Proceedings of the STI Berlin & CSW PhD Workshop*. Citeseer, 2008.

[16] D Vijaya Kumar and B Srinivasa Rao. Veracity finding from information provide on the web. *Computer Science & Telecommunications*, 28(5), 2010.

[17] Marianela Garcá Lozano, Ulrik Franke, Magnus Rosell, and Vladimir Vlassov. Towards automatic veracity assessment of open source information. In *2015 IEEE International Congress on Big Data*, pages 199–206. IEEE, 2015.

[18] Donghua Pan, Shaogang Qiu, and Dawei Yin. Web page content extraction method based on link density and statistic. In *2008 4th International Conference on Wireless Communications, Networking and Mobile Computing*, pages 1–4. IEEE, 2008.

[19] Xiaoxin Yin, Jiawei Han, and S Yu Philip. Truth discovery with multiple conflicting information providers on the web. *IEEE Transactions on Knowledge and Data Engineering*, 20(6):796–808, 2008.

[20] Fan Zhang, Li Yu, Xiangrui Cai, Ying Zhang, and Haiwei Zhang. Truth finding from multiple data sources by source confidence estimation. In *2015 12th Web Information System and Application Conference (WISA)*, pages 153–156. IEEE, 2015.

[21] Liang Zhao, Ting Hua, Chang-Tien Lu, and Ray Chen. A topic-focused trust model for twitter. *Computer Communications*, 76:1–11, 2016.