STM JOURNALS

Journal of Artificial Intelligence Research & Advances
ISSN: 2395-6720 (Online)
Volume 5, Issue 3
www.stmjournals.com

A New Evaluation Index for Application of Machine Learning Algorithms to Determine Trust in Skewed Social Media Data

Shifaa Basharat Fazili[1,*], Manzoor Ahmad[2]
[1]Research Scholar, Department of Computer Science, University of Kashmir, Hazratbal, Srinagar, Jammu and Kashmir, India
[2]Scientist D, Department of Computer Science, University of Kashmir, Hazratbal, Srinagar, Jammu and Kashmir, India

Abstract

In this study, we study the problem of accuracy paradox which arises on the application of machine learning algorithms for inference of trust in skewed social media data. Skewness is defined as the under-representation of one class over another in a binary classification problem. We achieved our purpose of identifying the accuracy paradox problem in various algorithms by identifying a new evaluation index called predictive index. The dataset used was that of Twitter, one of the most commonly used collaborative system, which has experienced enormous growth in a small duration of time. It has evolved from a microblogging service to a major news source used by people as a platform to share and disseminate information about current events. However, not all information posted on Twitter is trustworthy or useful in providing information about the event. Gossips, fake news, etc., are also a part of genuine news. The main aim of this study is to tackle the issue of accuracy paradox, a major problem when dealing with social media research, where the data extracted by us was highly skewed. This high skewness in the dataset gives us biased information about the performance of our machine learning algorithms.

Keywords: *Linear discriminant analysis, Gaussian Naive Bayes, Java script object notation, True reliable rate, True unreliable rate*

****Author for Correspondence*** E-mail: fazilishifaa@gmail.com

INTRODUCTION

Social media is quickly arising as a new and popular form of media. The last ten years have seen a huge rise in social media influence, not only in terms of communication but also in terms of business, travel, tourism, relationships, food, etc. Some social media services that are used by people include Facebook, Twitter, and LinkedIn. A study by the Pew Research Center has identified the internet as the most important resource for the news for people under the age of 30 in the US and the second most important overall after television [4]. While social media is mostly used for everyday chatter, it is also used to share news and other important information [5, 6]. Now more than ever, people turn to social media as their source of news [7–9]; this is especially true for breaking-news situations, where people crave rapid updates on developing events in real time. As Kwak et al. have shown, over 85% of all trending topics on Twitter are news [10]. Moreover, the ubiquity, accessibility, speed and ease-of-use of social media have made them invaluable sources of first-hand information. Twitter, for example, has proven to be very useful in emergency and disaster situations, particularly for response and recovery [10, 17].

The principal aim of social media networks is to provide a common platform for users to manipulate and share content. In order to encourage participation, most social media networking sites have no or minimal barriers to entry. Consequently, anyone can be the source of content. Such diversity of sources brings the trustworthiness of the content into question as social media is relied upon by millions of users as an information source [1]. Thus, considering every facet of the impact of social media on our lives, we realize the

factors that make it a great source of information dissemination also make it a fertile ground for the creation and spread of unsubstantiated and unverified information about events happening in the world.

In addition to low entry barrier, open platform and anonymity makes social media sites such as Twitter and Facebook vulnerable to activities of ill intentions such as deception. Deception is any distortion with an intention to mislead users, analysts, organizations, etc. A distortion can be about content, source, identity, age, sex, or location among many [2].

For example, journalists have erroneously considered Wikipedia content authoritative and reported false statements [3]. In one of the most recent incidents in the USA, Dow Jones index plunged 140 points due to a rumor Tweet posted from a news agency's (Associated Press) Twitter account [11]. The estimated temporary loss of market cap in the S&P 500 totaled $136.5 billion. The rumor mentioned that US President Barack Obama has been injured in twin explosions at the White House. In case of England Riots, online social media was responsible for spreading and instigating violence amongst people. Many rumors propagated during the riots, which resulted in large-scale panic and chaos among the citizens [12]. Given these vulnerabilities, the identification of trustworthy content/agent becomes apparent. The identification of trustworthy content is done by incorporating basic machine learning algorithms in the trust detection framework discussed in later sections of the study.

The study has been organized into a discussion on trust inference methodology for Twitter, the literature review, detailed data skewness and the accuracy paradox problem, following which we have discussed about the proposed evaluation index for analysis of the machine learning algorithms, experiments and results. Towards the end of this study, we have discussed conclusion and future scope.

TRUST INFERENCE METHODOLOGY

The trust inference framework can be divided into a number of parts as shown below (Fig. 1). It consists of the following components [14]:

1. Twitter data extraction
2. Data analysis and interpretation
3. Program-based annotation
4. Learning
5. Results

Twitter Data Extraction: The extraction of data from Twitter involved creating an application on Twitter which is followed by using our OAuth credentials for data extraction. The output of data extraction step is a JSON file that consists of the entire tweet metadata.

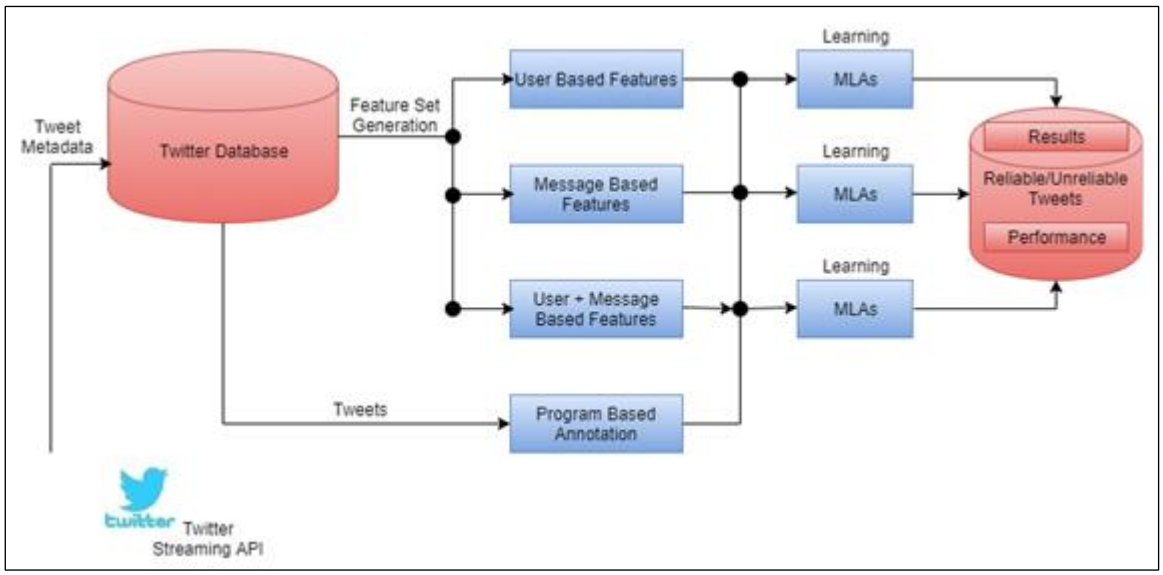

Fig. 1: *Trust Inference Framework.*

Data Analysis and Interpretation: This step deals with the generation of features for use in learning. Several constraints were placed on the analysis of raw JSON file. In particular, we focused on only English tweets and extracted both the user-based and message-based features from the file through a series of Java and Python programs as shown in Fig. 2. When extracting the message-based features, JTweet analyzes the entire tweet taking care of swear words and the internet slang.

Program Based Annotation: JAnnotate was used to prepare annotated data for use in the training. Input to JAnnotate program was a set of features as mentioned in Table 1 that were selected based on people's feedback, who were asked to indicate the features that they typically paid attention to while reading a particular tweet and rate the features on a 5-point Likert scale based on the features degree of impact on the reliability of a tweet [7].

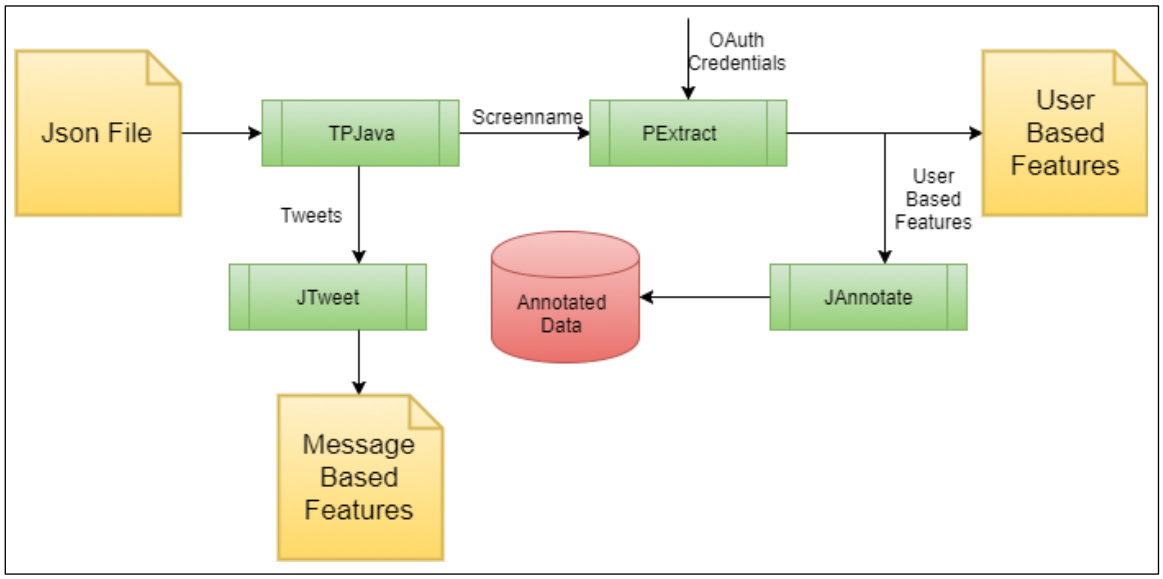

Fig. 2: *Twitter Data Analysis and Interpretation.*

Table 1: *Features Used for Annotation.*

Attribute Name	Impact on Credibility Score
Status_count	Status_count>100 and Status_count <1000, score++
	Status_count>1000 and Status_count <10000, score+2
	Status_count>10000, score+3
Favorite_count	favorite_count>0 , score++
Verified?	Verified=0,score—
	Verified=1,score++
Default_profile_image?	Default_profile_image=0,score++
	Default_profile_image=1,score—
Registration age	Registration age>=5,score++
	Registration age<1,score—
Follower/friend	Follower/friend>0 and follower/friend<1,score—
	Follower/friend>1 and follower/friend<10,score++
	Follower/friend>10 and follower/friend<100,score+2
	Follower/friend>100,score+3
Listed_countt	Listed_count>0 and listed_count<100,score++
	Listed_count>100,score+2
Retweet_count	Retweet_count>=5,score++

Table 2: *Twitter Features for Reliability Check.*

Attribute Name	Attribute Type	Attribute Information
Friends_count	Continuous/ user based	Number of people a twitter user follows
Follower friend ratio	Continuous/user based	Ratio of number of people a twitter user is followed by and the friends count
Status_count	Continuous/user based	Number of tweets a user has posted on his timeline
Retweet_count	Continuous/user based	Number of times the users tweet has been shared
Age	Continuous/user based	Registration age of a user
Favorite_count	Continuous/user based	The number of Tweets the user has liked in the account's lifetime.
Listed_count	Continuous/user based	The number of public lists that the user is a member of
Location	Categorical/user based	Whether the user has mentioned his location in his profile or not
Geo_enabled	Categorical/user based	Whether the user has enabled the possibility of geo tagging or not
Verified	Categorical/user based	Whether the user has a verified badge
Dp	Categorical/user based	Whether the user has altered the theme and background of his twitter profile or not
Dpi	Categorical/user based	Whether the user has uploaded a profile image, or a normal egg avatar has been used
Protected	Categorical/user based	Whether the user is protected or not. If the user is protected, then his tweet is visible only to his friends.
Sentiment	Ordinal/message based	Whether the tweet is positive, negative or neutral
No_of_chars	Continuous/message based	Total number of characters in the tweet (max chars=140)
No_of_words	Continuous/message based	Total number of words in a tweet
Pronoun_count	Continuous/message based	Number of pronouns in a tweet
Retweet	Categorical/message based	Whether a tweet is a retweet or not
url_count	Continuous/message based	Number of URLS in a tweet
Swearword_count	Continuous/message based	The number of swear words a tweet consists of
Special_symbols	Continuous/message based	The number of special symbols a tweet has like @, #, $,
Class	Categorical	Whether the tweet is reliable/ not reliable

Learning: Learning involves applying basic supervised machine learning algorithms used for classification. We have used five basic machine learning algorithms on various Twitter features (as mentioned in Table 2) and compared them based on their accuracy and predictive power. The algorithms used include Gaussian Naïve Bayes, Linear Discriminant Analysis, Logistic Regression, linear Regression and Support Vector Machines.

Results: Our Trust Inference Framework gives us four different types of outputs namely:

- True Reliable: The Tweets that have been identified as reliable and are actually reliable.
- False Reliable: The Tweets that have been identified as reliable but are actually unreliable.
- True Unreliable: The Tweets that have been identified as unreliable and are actually unreliable.
- False Unreliable: The Tweets that have been identified as unreliable but are actually reliable.

The output obtained is then used for the analysis of various algorithms by using the proposed evaluation index, namely predictive index.

RELATED WORK

Though Twitter acts as a real-time news source with people acting as sensors and sending event updates from all over the world, rumors spread via Twitter have been noted to cause considerable damage [18]. Thus, one major challenge in utilizing Twitter content is

to extract quality content from the huge amount of data posted on it. Different researchers have used different computational strategies such as classification or ranking one of the common machine learning techniques to study Twitters trust identification problem. Zhao et al. in their paper have developed a Twitter-LDA model (unsupervised topic modeling) for comparing the news topics from Twitter and a traditional news dissemination medium (*The New York Times*) [19]. They compared a total of 1,225,851 tweets (news articles) with 11,924 news articles from *The New York Times* and the results obtained reveal that Twitter can be a good source of entity-oriented topics that have low coverage in traditional media and although Twitter users show relatively low interest in world news, they actively spread news about important world events. Castillo et al. in their work [20] assessed the credibility of news propagated through tweets using a number of machine-learning schemes including SVM, decision trees, decision rules and Bayes network. The results obtained showed that J48 decision tree algorithm outperformed all the other algorithms, achieving an accuracy of 89%. They evaluated their results with respect to data annotated by humans as ground truth. A different approach adopted by Gupta et al. was to determine the credibility of events on Twitter. They proposed a credibility analysis approach enhanced with event graph-based optimization that resulted in an accuracy of ~86% compared to ~72% of the decision tree approach [21]. Another type of research work carried out by Ratkiewicz et al. [22] in this area was the development of *Truthy,* a live web service built to study information diffusion on Twitter and compute trustworthiness score for a public stream of microblogging updates related to an event to detect political smears, astroturfing and other forms of social pollution. Mendoza et al. in their work explored the attributes of rumors and true news and their results revealed that the propagation of tweets related to rumors versus true news differed and could be used to develop automated classification solutions to identify correct information [23]. They also concluded that the tweets related to rumors contained more questions versus news tweets. Barbier and Liu [24] in their work proposed a method to find provenance paths leading to sources of the information to evaluate its credibility. Guha et al. [25] studied the problem of propagating trust and distrust among Epinions users who assign positive and negative ratings to one another and thereby use trust information to rank users and using that rank the content generated by the users.

Saikaew and Noyunsan [26] have discussed two different classes of credibility computation in social media, namely, web-page-independent and webpage-dependent where web-page-independent uses messages in social media for computing credibility by comparing the messages with trusted news sources which if similar then the credibility score of the message is high and webpage-dependent approaches use features of each social media for computing credibility such as like, comment, and re-tweet. In this study, we have focused on webpage-dependent approaches.

DATA SKEWNESS AND ACCURACY PARADOX

The most commonly used evaluation metric for machine learning algorithms, accuracy, suffers from a paradox such that the models with lower levels of accuracy may have better predictive power than models with higher levels of accuracy. Such a condition is referred to as accuracy paradox. For example, consider a dataset with more than 90% of the observations belonging to one class and the remaining observations belonging to other class, then a machine learning algorithm that classifies all the examples into the majority class will give an accuracy of more than 90% thereby misleading us about the quality of the algorithm being very good although it has 0% predictive power. Although the classification error rate is optimized, the high accuracy models fail to capture crucial information transfer in the classification task [13].

This problem of accuracy paradox in machine learning algorithms arises due to the skewness in the data for which the models are trained. Data skewness refers to the over representation of one class over another. Skewed data is a challenging problem for classification, in machine learning, as research shows that the

algorithms trained with un-skewed datasets outperform those with skewed datasets [15, 16, 24]. Using skewed datasets for classification result in models that are biased towards majority labels and thereby give us misleading results.

Motivated by this, we developed a new evaluation index known as the predictive index for measuring the classification performance that takes into consideration the degree of skewness in the dataset.

Predictive Index

After implementing the machine learning framework for our problem, one needs to find out the effectiveness of the model. Among a majority of evaluation metrics for machine learning algorithms, the most common metric is that of accuracy. However, accuracy is a good measure when the target variable classes in the data are nearly balanced which is not the case, we are implementing as the data used in this study is highly imbalanced with 89% of the tweets belonging to reliable class and only 11% tweets belonging to the unreliable class. Thus, a new index has been proposed in this study known as the predictive index which is used to determine the predictive power of an algorithm. Predictive power can be defined as the ability of an algorithm to make correct predictions from both the positive and negative classes in a binary classification problem. Predictive index (P_i) can be calculated by the following formula:

$$P_i = 1 - \left| \frac{tr(n_{ur}) - tur(n_r)}{n_r \bullet n_{ur}} \right| \quad (1)$$

Where,

P_i is the predictive index of an algorithm

tr refers to the true reliable tweets

tur refers to true unreliable tweets

n_r refers to the total number of reliable tweets

n_{ur} refers to the total number of unreliable tweets

Greater the value of predictive index, greater is the predictive power.

Experimental Procedure

In order to test the efficiency of our evaluation metric, we collected data from Twitters streaming API. A total of approximately 14K tweets were collected, the details of which are provided in Table 3.

Table 3: *Twitter Data Statistics.*

Event	Hashtags	Number of Tweets
Paris Attacks – 2015	#ParisAttacks, #prayforparis, #pray4Paris, #peaceforParis	548
Hamas Attacks - 2016	#Israel, #Gaza, #Hamas, #ISIS, #Palestine	2241
Random Tweets	#news, #TechNews, #NYTimes	11103

From a total of 14K tweets, we created four different combinations of training and test datasets (Table 4) to see how our proposed predictive index performs in determining the problem of accuracy paradox for skewed and un-skewed datasets.

Table 4: *Data Decomposition.*

	Training Data (No. of Tweets)	Test Data (No. of Tweets)
Skewed training and Test Data (STT)	7889R+896UR	4300R+896UR
Un-skewed Training and Test Data (UTT)	1000R+1000UR	792R+792UR
Skewed Training Un-Skewed Test Data (STUT)	11293R+896UR	896R+896UR
Un-skewed Training Skewed Test Data (UTST)	1000R+1000 UR	11189R+792UR

**R stands for reliable and UR stands for unreliable*

RESULTS

The results obtained for all the four data decompositions have been listed in the tables 5 below. From the results we observe that the problem of accuracy paradox is clearly visible when the training data is skewed irrespective of whether the test data is skewed or not. In Table 5, the case of skewed training and test data accuracy paradox can be seen in linear discriminant analysis (LDA) and logistic regression which show an accuracy of ~83% but the true unreliable rate (TUR) is as low as 0.02, thereby implying that both the algorithms have identified almost all the tweets in the test data as reliable which renders these algorithms as unusable as they have a predictive index of almost 0 which implies zero predictive power.

Table 5: *Results for Skewed Training and Test Data.*

	Accuracy	**TRR**	**TUR**	**Predictive Index**
Naive Bayes	0.548884	0.463953	0.956473	0.50748
LDA	**0.829677**	**0.997907**	**0.022321**	**0.024414**
Logistic Regression	**0.828155**	**0.999535**	**0.021882**	**0.22347**
Linear Regression	0.17244	0	1	0
SVM	0.465743	0.418372	0.69308	0.72529

Similarly, in Table 6, the case of skewed training and balanced test data, i.e., unbalanced training data, accuracy paradox problem is clearly visible in LDA and logistic regression.

Table 6: *Results for Skewed Training and Un-skewed Test Data.*

	Accuracy	**TRR**	**TUR**	**Predictive Index**
Naive Bayes	0.736049	0.516741	0.955357	0.561384
LDA	0.507254	0.998884	0.015625	0.016741
Logistic Regression	0.499441	1	0	0
Linear Regression	0.5	0	1	0
SVM	0.533482	0.407366	0.659598	0.747768

Table 7: *Results for Un-skewed Training and Skewed Test Data.*

	Accuracy	**TRR**	**TUR**	**Predictive Index**
Naive Bayes	0.454803	0.419251	0.957071	0.46218
LDA	0.633837	0.624721	0.762626	0.862095
Logistic Regression	0.516818	0.495397	0.819444	0.675953
Linear Regression	0.066105	0	1	0
SVM	0.41282	0.398516	0.614899	0.783617

Table 8: *Results for Un-skewed Training and Test Data.*

	Accuracy	**TRR**	**TUR**	**Predictive Index**
Naive Bayes	0.679293	0.401515	0.957071	0.444444
LDA	0.779672	0.796717	0.762626	0.965909
Logistic Regression	0.790404	0.757576	0.823232	0.934344
Linear Regression	0.5	0	1	0
SVM	0.550505	0.103535	0.997475	0.10606

However, apart from linear regression that performs poorly in all the four cases of data decomposition with zero predictive index, accuracy paradox is not a problem for the other algorithms in case of balanced or un-skewed training data as can be seen in Tables 7 and 8.

CONCLUSION AND FUTURE SCOPE

Not only has the internet replaced traditional media as a source for obtaining news but it has also provided platform for common people to express and share their opinions. However, users online lack the clues that they have in the real world to access the reliability of information to which they are exposed, especially in case of inexperienced users, who can be easily mislead by false content on the social media. Although machine learning has been implemented to solve this problem of differentiating reliable and unreliable content, but basic machine learning algorithms fail to perform in case of skewed social media datasets.

In this study, we have used Twitter as the medium of information and shown that a fully automated reliability assessment framework based on basic supervised machine learning algorithms does not perform well in case of skewed social media data with more that 80% of the tweets belonging to reliable class and the remaining tweets being unreliable.

To this end, we proposed a predictive index that can be used to identify misleading accuracies of various algorithms. As part of future work, we plan to extend these experiments to larger datasets and also devise a machine learning algorithm based on hybrid methodology [27] that performs well irrespective of the skewness in the social media datasets.

ABBREVIATIONS

LDA: Linear Discriminant Analysis
GNB: Gaussian Naive Bayes
JSON: Java Script Object Notation
TRR: True Reliable Rate
TUR: True Unreliable Rate

REFERENCES

1. West AG, Chang J, Venkatasubramanian KK et al. Trust in collaborative web

applications. *Future Generation Computer Systems.* 2012; 28: 1238–1251p.
2. Huan Liu, Jiawei Han, Hiroshi Motoda. Uncovering deception. *Springer Journal Social Network Analysis and Mining.*
3. Pogatchnik S. Student hoaxes world's media on Wikipedia. http://www.msnbc.msn.com/id/30699302/.
4. TPR Center. Internet overtakes newspapers as news outlet. December 2008. http://pewresearch.org/pubs/1066/internet-overtakes-newspapers-as-newssource [pewresearch.org; posted 23-December-2008].
5. Java A, Song X, Finin T et al. Why we twitter: Understanding microblogging usage and communities. In Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis. *ACM*. 2007; 56–65p.
6. Naaman M, Boase J, Lai C-H. Is it really about me? Message content in social awareness streams. *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*. *ACM*. 2010; 189–192p.
7. Laird S. How social media is taking over the news industry. April 2012. http://mashable.com/2012/04/18/social-media-and-the-news/mashable.com; posted 18-April-2012.
8. Kwak H, Lee C, Park H et al. What is twitter, a social network or a news media? In: Proceedings of the 19th international conference on World Wide Web. *ACM*. 2010; 591–600p.
9. Stassen W. Your news in 140 characters: Exploring the role of social media in journalism. *Global Media Journal*. 2010; 4(1): 116–131p.
10. CNBC. 2013. False rumor of explosion at white house causes stocks to briefly plunge; ape confirms its twitter feed was hacked. http://www.cnbc.com/id/100646197.
11. Valverde-Albacete, Francisco J, Carmen Peláez-Moreno. 100% classification accuracy considered harmful: The normalized information transfer factor explains the accuracy paradox. *PloS One* 2014; 9.1: e84217.
12. Shifaa B, Ahmad M. Inferring trust from message features using linear regression and support vector machines. International *Conference on Next Generation Computing Technologies.* Springer, Singapore. 2017.
13. Richards J, Lewis P. 2011. How Twitter was used to spread and knock down rumours during the riots. http://www.guardian.co.uk/uk/2011/dec/07/how-twitter-spread-rumours- riots.
14. Zeldin W. Venezuela: 2010. Twitter users arrested on charges of spreading rumors. http://www.loc.gov/lawweb/servlet/lloc news?disp3 l205402106 text.
15. Soroush V. Automatic detection and verification of rumors on Twitter. *Diss. Massachusetts Institute of Technology* 2015.
16. Gupta A, Kumaraguru P. Credibility ranking of tweets during high impact events. Proceedings of the 1st workshop on privacy and security in online social media. *ACM*. 2012.
17. Zhao WX, Jiang J, Weng J et al. Comparing twitter and traditional media using topic models. In: Proceedings of the 33rd European conference on Advances in information retrieval (Berlin, Heidelberg, 2011), ECIR'11. *Springer-Verlag*. 338–349p.
18. Castillo C, Mendoza M, Poblete B. Information credibility on twitter. In: Proceedings of the 20th international conference on World Wide Web (New York, NY, USA, 2011). WWW '11. *ACM*. 675–684.
19. Gupta M, Zhao P, Han J. Evaluating event credibility on twitter. In: Proceedings of the 2012 SIAM International Conference on Data Mining. *Society for Industrial and Applied Mathematics.* April 2012; 153–164.
20. Ratkiewicz J et al. Truthy: Mapping the spread of AstroTurf in microblog streams. Proceedings of the 20th international conference companion on World Wide Web. *ACM*. 2011.
21. Mendoza M, Poblete B, Castillo C. Twitter under crisis: Can we trust what we write? In: Proceedings of the first workshop on social media analytics (New

York, NY, USA, 2010), SOMA '10. *ACM*. 71–79p.

22. Barbier G, Liu H. Information provenance in social media. Social Computing, Behavioral-Cultural Modeling and Prediction. 2011; 276–283p.
23. Guha R, Kumar R, Raghavan P et al. Propagation of trust and distrust. In: Proceedings of the 13th international conference on World Wide Web. *ACM*. 2004; 403–412p.
24. Saikaew KR, Noyunsan C. Features for measuring credibility on Facebook information. *International Scholarly and Scientific Research & Innovation*. 2015; 9.1: 174–177p.
25. Basharat S, Chachoo M. On Linear vs. Hybrid Configuration: An Empirical Study. In: *Commune*. 2015; 180–184p.

Cite this Article
Shifaa Basharat Fazili, Manzoor Ahmad. A New Evaluation Index for Application of Machine Learning Algorithms to Determine Trust in Skewed Social Media Data. *Journal of Artificial Intelligence Research & Advances*. 2018; 5(3): 49–57p.