Journal of Advanced Zoology

ISSN: 0253-7214
*Volume **44** Issue **S-6 Year 2023** Page **608:615***

System and Method for Truth Discovery in social media Big Data

Nirupama KS[1], Prof. Akram Pasha[2], Greeshma Reddy[3], Sree Chandana[3], Roopa[4]

[1,2,3,4]C&IT REVA UNIVERSITY
Email: r17cs270@cit.reva.edu.in[1], akrampasha@reva.edu.in[2], r17cs251@cit.reva.edu.in[3], r17cs505@cit.reva.edu.in[4]

**Corresponding author's E-mail: r17cs284@cit.reva.edu.in*

Article History

Received: 06 June 2023
Revised: 05 Sept 2023
Accepted: 29 Nov 2023

Abstract

Within the span of enormous information and the coming of numerous advancements in the communication technologies, at every tick of the clock, enormous sums of information is produced from different sources. One such source of data generation is social media. However, such data carries much of the noisy, uncertain, and untrustworthy data. In this way, finding independable information from loud information is one of the characteristic challenges of huge information focusing on the esteem characteristic of enormous information. Therefore, in this article, an attempt is made to target a few challenges arriving from "misinformation spread", "data sparsity" or the "long-tail wonder" in the domain of social media data analytics. The study uses an instance from the Online Social Network (OSN) datasets to develop scalable to wide-range social sensing by consolidating Scalable Robust Trust Discovery (SRTD) plots to address the mentioned challenges utilizing the distributed parallel computing framework. The dataset picked for investigation includes 128,483 tweets which incorporates 20% deception, 80% retweets bringing about 0.05 milliseconds utilizing Spark parallel processing.

Keywords: *Spark parallel processing; WorkQueue; SRTD.*

1. Introduction

In the work of [8], Identifying deception on social media is an amazingly critical but too challenging issue. Web pages play a vital part in combating misinformation, but they stand in need of a master examination which restrains an opportune reaction. Web-based media moreover authorizes the wide inciting of "misinformation", that is, news with purposely wrong data. Misinformation by means of online media can have basic negative social impacts. Recognizing and moderating fake news also presents special difficulties [7]. We characterize news as any story or ensure with an explanation in it and conversation within the light of the reality that the social wonders of a news story or ensure spreading or diffusing through the Twitter network. That's, rumors are naturally social and incorporate the sharing of cases between people [6]. Compared with conventional media, data on the web is regularly distributed quickly, but with less guaranteed quality and validity. Whereas clashing information is taken note exceptionally frequently on the net, common clients still believe Web data [5]. The broad increase in untrue news has the potential for exceptionally negative affect on people and society. In this manner, the discovery of untrue news on Social media has as of late ended up an emerging research that's drawing in striking attention [2]. Individuals depend on the written surveys in. decision-making forms, for the determination, handling of items and administrations positive/negative surveys encouraging/discouraging surveys are used. Composed reviews offer assistance to improve the quality of items and administrations for benefit suppliers. These surveys have become a significant thing, almost a victory of a trade, for positive surveys bring benefits for an enterprise, negative reviews can possibly affect validity and cause financial losses [3].

Problem Formulation

In this portion we work out the truth disclosure issue in enormous information social media. Here we take X number of sources equal to A1, A2, A3.....AX and with Y number of observations(claims) equal to B1, B2, B3....BY. Consider A_i denotes i-th source and Bj denotes j-th claim and $R^k_{i,j}$ is the result we get by source A_i and B_j at time k.

As per the data set we took from twitter, each person's account is considered as a source and claim is a statement or report provided by the sources.

Now we define B_j = True and B_j = False for claim and every claim has truth label z^*_j where z_j=1 when B_j is True and z_j =0 when B_j is False. In order to find the claim truth, we estimate truthfulness for each claim and reliability of source.

Truthfulness of claim: It is about agreeing whether the claim is True or False. Consider D_j truthfulness to claim $\mathbf{B_j}$. The more D_j the more $\mathbf{B_j}$ will be true. So, we conclude:

Result---->B_j = True ***(1)***

Reliability of source: Defines how trustworthy the source is.

Reliability R_i for source A_i:

Result---->B_j = True | $AB_{i,j}$ = True ***(2)***

Where source A_i reporting claim B_j is true

Credibility score *(C):* It provides the report which contributes extra evidence for truthfulness of claim.

Credibility score incorporates properties like: Independent score, attitude score, and uncertainty score.

- **Independent Score ($\pi^k_{i,j}$):** Score ranges from 0 to 1 measures the report has been forwarded, copied or independent. Higher score is provided to the independent claim which is not dependent on any claim..
- **Attitude Score ($\alpha^k_{i,j}$):** Scores include 1, -1, 0 representing the attitude score for claiming true, false or means nothing.
- **Uncertainty Score ($\beta^k_{i,j}$):** Score ranges from 0 to 1 which measures the uncertainty of the report.

So finally defining the Credibility Score from the above terms from the report of source A_i on claim B_j at time k :

$C^k_{i,j} = \alpha \times (1-\beta) \times \pi$ ***(3)***

From above definition we can understand

- Claim is true or false.
- The confidence in the claim.
- Claim is dependent or independent.

For X source and Y claims we characterize a matrix called Time Series Matrix ($TSM_{X \times Y}$) where $C^k_{i,j}$ speaks for the Credibility Score during time k.

$TSM_{i,j}$ ={ $C_{i,j}{}^1$, $C_{i,j}{}^2$, $C_{i,j}{}^3$ .....$C_{i,j}{}^k$} ***(4)***

TSM is provided as input from the data we have chosen from social media and estimates Dj as the output.

Table 1. Terms described from above equations.

A_j	ith source
B_j	jth claim
$C^k_{i,j}$	Credibility Score
z^*_j	Truth label of jth claim
R_i	Reliability of source A_i
$CS_{i,j}{}^k$	Contribution Score
K	Size of sequence

R_i^{K+1-k}	Weight provided based on the report whether it is the latest or old one.

Computing: $\forall j\ 1 \leq j \leq Y$ ***(5)***

Result----> B_j = True | TSM

Contribution Score (CS): With the help of TSM matrix, we combine the credibility scores of all the reports and define what is Contribution Score.

Considering the reliability and credibility Score from report of any source

- The CS score is high for reliable sources.
- Independent claim reports get the highest CS score.
- Less uncertainty reports get a high CS score.

$$CS_{i,j} = sgn(C_{i,j}^k)\Sigma^K{}_{k=1} R_i^{K+1-k}|C_{i,j}^k| \quad \textit{(6)}$$

IV. ALGORITHM

Joining the truthfulness of claims and reliability of source we compute:

$$Ri = \frac{\Sigma_{j\in F(i)}|CS_{i,j}|(\theta(CS_{i,j})D_j+(1-\theta(CS_{i,j}))(1-D_j))}{\Sigma_{j\in F(i)}|CS_{i,j}|} \quad \textit{(7)}$$

Where $\theta(x) = \begin{cases} 1, & x > 0 \\ 0, & x \leq 0 \end{cases}$

Here F(i) may be a set of perceptions (claims determined by source Ai).

- To find the claim truthfulness, we update the score of truthfulness.

$$CT_j = \Sigma_{i\in K(j)}|CS_{i,j}| \quad \textit{(8)}$$

Where $CS_{i,j}$ is a contribution score.

$$D_j = \frac{1}{1-\exp(-CT_j)} \quad \textit{(9)}$$

$K_{(j)}$ are all those sources who claim C_j.

For scalable algorithms we make sure to divide the input TSM into some variable submatrices (V) where every submatrix includes a subset of sources in A and claims reported by that set of sources.

Now calculate the reliability for each source and truthfulness of claim for each substance.

- We find the partial truthfulness claim, where the score has TC_1^v, TC_2^v,.. so on.
- Sum all the partial scores of every submatrix to get the final score.
- Using the sigmoid function normalizes the score.
- CS score is updated.

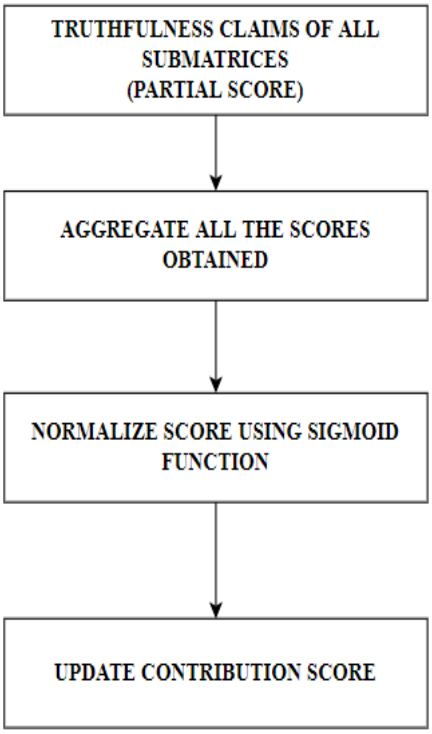

Fig-1: Steps to update CS score

Finally, $CT_j = \Sigma\, CT_j^v$

$$CT_j^v = \Sigma\, CS_{i\,j} \qquad \textit{(10)}$$

From (9),

$$D_j = \frac{1}{1 + \exp(-CT_j)} \qquad \textit{(11)}$$

ALGORITHM: Truth Discovery (SRTD)

Input: Matrix (TSM)

Output: Truthfulness of claim

Step 1: Set $R_i = 0.5$ for every $i \leq X$

Step 2: Set credibility score attributes which are attitude score, uncertainty score, independent score.

Step 3: Iteration = 100

Step 4: Divide TSM into V submatrices, A(v) be the same source in each V submatrix.

~~**while** Dj until it reaches maximum iteration **do**~~

foreach v, $1 \leq v \leq j$ **do**

foreach i, $1 \leq i \leq A(v)$ **do**

foreach j, $1 \leq i \leq y$ **do**

if $TSM_{i,j}$ is true **then**

calculate contribution score using ***(6)***

end

end

end

foreach i, $1 \leq i \leq A(v)$ **do**

calculate Ri using ***(7)***

end

foreach j, -1≤j≤y **do**

calculate CTjv using ***(10)***

end

calculate Dj using ***(11)***

end

end

foreach j, 1≤j≤y **do**

if Dj ≥ threshold **then**

Zj* =1

else

Zj* =0

end

end

IMPLEMENTATION

HTCondor for work queue: HTCondor is a sophisticated workload management framework for jobs that need a lot of computing power. HTCondor has a work queueing process, scheduling strategy, priority scheme, resource control, and resource management, much like most full-featured batch programs. Users upload serial or concurrent jobs to HTCondor, which puts them in a queue and decides when and where to execute them based on a policy.

Spark for parallel processing: All that you will do in Apache Spark is to peruse some information from a source and burden it into Spark. You will then process the information and hold the halfway outcomes, lastly composing the outcomes back to an objective. In any case, in this process, you need an information design to hold the information in Spark.

Fig-2: Flow chart for updating contribution score

Data collection:

Genuine world information set being collected from twitter called Dallas shooting that happened on July seventh 2016. A bunch of police officers were trapped in Dallas murdering around five officers

and nine others were extremely harmed. Incident took place for 8days at Dallas, USA. The dataset contributes 128,483 tweets where 1.17 tweets belong to each user.

In this data set 1.4 percent of sources claim more than two claims whereas 90 percent of sources contribute as if it were one claim. 20 percent was found to be misinformation in the dataset and 80 percent of them are retweeted.

Pre-processing:

The steps included in pre-processing are:

- Cluster comparable tweets into the same cluster to create claims.
- Derived credibility score
- Generate the TSM matrix
- Generate truth labels

Evaluation:

Check for the efficiency of SRTD over the HTConder and Spark Parallel processing and compare the efficiency of both.

3. Results and Discussion
Experimental setup:

Operating system: Windows 7,8,10 Ultimate, Linux, Mac.

Front-End **:** Python.

Coding Language **:** Python.

Software Environment**:** Anaconda (jupyter or spyder).

Discussion of Results:

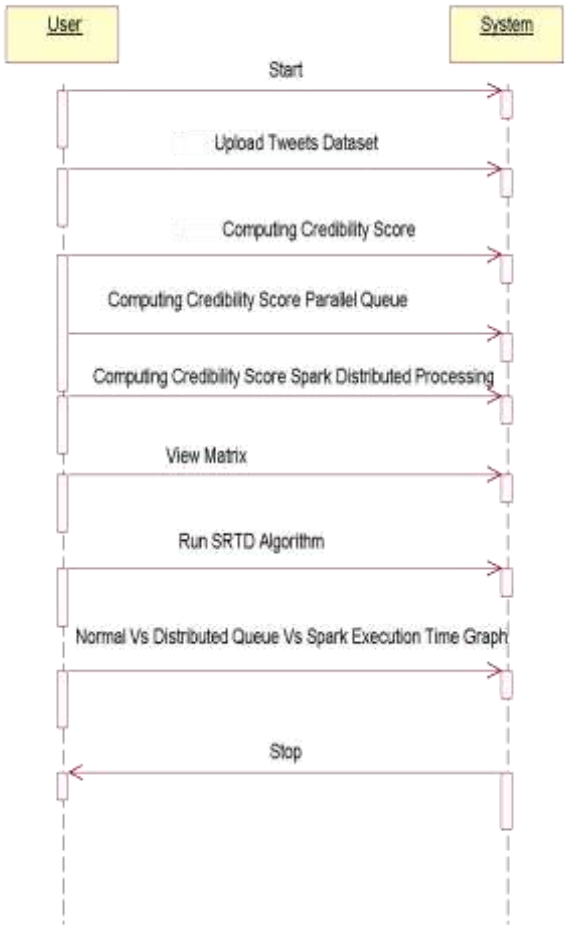

Fig-3: Sequence Diagram

The dataset being gathered from the twitter called Dallas shooting dataset occurred on July 7,2016 is being uploaded when the user is approached to upload the dataset followed by checking the validity score, we will actually want to compute attitude score, uncertainty score and independent score. As we have effectively referenced the significance of figuring these scores. At last, we gauge the productivity of execution of SRTD calculation over HTCondor workqueue according to the past creator which brings about 0.058 milliseconds yet we have expanded utilizing Spark parallel processing bringing about 0.005 milliseconds which is similarly lesser than past creator work.

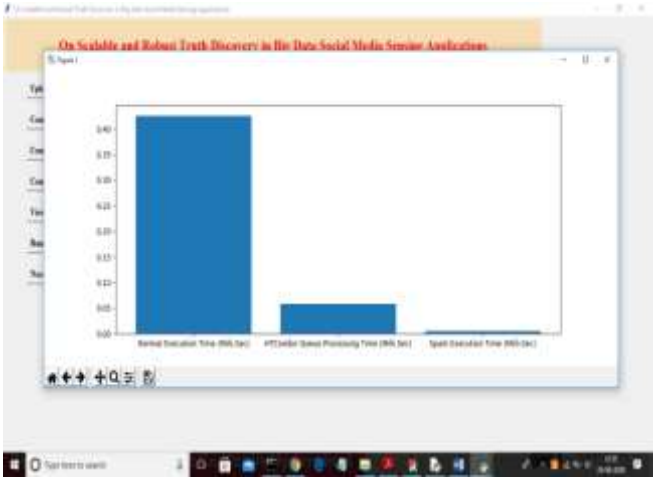

Fig-4: Comparison on execution time for Normal queue processing, HTCondor queue processing and Spark processing

4. Conclusion

In this paper, we evaluated existing SRTD algorithms over Spark parallel preprocessing and HTCondor work queue in order to compare the efficiency and finally resulted as Spark parallel preprocessing has more efficiency compared to HTCondor work queue. In our arrangement, we expressly see the supply reliability, record validity, also a source's notable ways to accurately handle the deception unfold along with record insufficiency challenges within the truth discovery issue. Apache Spark allows the system to perform multidimensional activities such as processing, questioning and producing analytics at high speeds and looking long-term, it appears likely that Apache Spark is getting to be the foremost well-known stage for big data. A vital figure in this context is Apache Start is an open-source system which increments its request in an something else costly exclusive innovation advertisement. Apache Spark is seen as a competitor or successor to MapReduce. There are a few specialists who still consider Spark system at its incipient stages and it can right presently support as it were one or two of operational analytics. With the progression of innovation, advanced news is more broadly uncovered to users globally and contributes to the increase of spreading deceptions and disinformation online. Fake news can be found through prevalent stages such as social media and the Web. In any case, fake news intends to persuade the reader to accept wrong data which considers these articles troublesome to see. The rate of creating computerized news is expansive and fast, running day by day at each moment, in this way it is challenging for machine learning to viably identify fake news.

References:

1. Tschiatschek S, Singla A, Gomez Rodriguez M, Merchant A, Krause A. Fake news detection in social networks via crowd signals. InCompanion Proceedings of The Web Conference 2018 2018 Apr 23 (pp. 517-524).
2. Shu K, Bernard HR, Liu H. Studying fake news via network analysis: detection and mitigation. InEmerging Research Challenges and Opportunities in Computational Social Network Analysis and Mining 2019 (pp. 43-65). Springer, Cham.
3. Ruchansky N, Seo S, Liu Y. Csi: A hybrid deep model for fake news detection. InProceedings of the 2017 ACM on Conference on Information and Knowledge Management 2017 Nov 6 (pp. 797-806).
4. Bruns A. Big social data approaches in Internet studies: The case of Twitter. Second international handbook of Internet research. 2020:65-81.
5. Ahmed O, Gupta S, Hasibuddin M. Truth Discovery in Big Data Social Media Sensing Applications.
6. Zubiaga A, Liakata M, Procter R, Wong Sak Hoi G, Tolmie P. Analysing how people orient to and spread rumours in social media by looking at conversational threads. PloS one. 2016 Mar 4;11(3):e0150989.
7. Al-Garadi MA, Hussain MR, Khan N, Murtaza G, Nweke HF, Ali I, Mujtaba G, Chiroma H, Khattak HA, Gani A. Predicting cyberbullying on social media in the big data era using machine learning algorithms: review of literature and open challenges. IEEE Access. 2019 May 22;7:70701-18.

8. Paul S, Joy JI, Sarker S, Ahmed S, Das AK. Fake News Detection in Social Media using Blockchain. In2019 7th International Conference on Smart Computing & Communications (ICSCC) 2019 Jun 28 (pp. 1-5). IEEE.
9. Iftikhar R, Khan MS. Social Media Big Data Analytics for Demand Forecasting: Development and Case Implementation of an Innovative Framework. Journal of Global Information Management (JGIM). 2020 Jan 1;28(1):103-20.
10. Shu K, Sliva A, Wang S, Tang J, Liu H. Fake news detection on social media: A data mining perspective. ACM SIGKDD explorations newsletter. 2017 Sep 1;19(1):22-36.
11. Kumar, Akshi, and Saurabh Raj Sangwan. "Rumor detection using machine learning techniques on social media." International Conference on Innovative Computing and Communications. Springer, Singapore, 2019.
12. Zhang DY, Wang D, Zhang Y. Constraint-aware dynamic truth discovery in big data social media sensing. In2017 IEEE International Conference on Big Data (Big Data) 2017 Dec 11 (pp. 57-66). IEEE.
13. Li Y, Gao J, Meng C, Li Q, Su L, Zhao B, Fan W, Han J. A survey on truth discovery. ACM Sigkdd Explorations Newsletter. 2016 Feb 25;17(2):1-6.
14. Nigade M, Raut M, Mane P, Phadatare S. Truth Discovery in Big Data Social Media Application. © Journal of Data Mining and Knowledge Engineering. 2019:40-4.
15. Singleton J, Booth R. An axiomatic approach to truth discovery.
16. Thiyagaraj, Mr P. Bastin, and A. Aloysius. "A survey on truth discovery methods for big data." International Journal of Computational Intelligence Research 13.7 (2017): 1799-1810.
17. Singh, Mahima, and Nalini Sampath. "Truth Discovery Algorithm on Big Data Social Media Tweets." 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA). IEEE, 2020.
18. Wang, Dong, Jermaine Marshall, and Chao Huang. "Theme-relevant truth discovery on twitter: An estimation theoretic approach." Proceedings of the International AAAI Conference on Web and Social Media. Vol. 10. No. 1. 2016.
19. Zhang, Daniel Yue, et al. "On robust truth discovery in sparse social media sensing." 2016 IEEE International Conference on Big Data (Big Data). IEEE, 2016.
20. Zhang, Daniel Yue, et al. "Towards reliable missing truth discovery in online social media sensing applications." *2018 IEEE/ACM* International Conference on Advances in Social Networks Analysis and Mining *(ASONAM)*. IEEE, 2018.