International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 05 Issue: 05 | May-2021 ISSN: 2582-3930

Truth Discovery Efficiency Prediction of Social Media Tweets

Riya Goel[1], Yukta Taneja[2], Ms. Sonam Sharma[3]

[1]Riya Goel, Department of Computer Science and & Engineering, SRMIST
[2]Yukta Taneja,Department of Computer Science and & Engineering, SRMIST
[3]Sonam Sharma, Assistant Professor, SRMIST

Abstract -In this work, an approach is made for the detection of truth in social media tweets in big data concepts, based on a truth discovery algorithm. The attempt is to mainly improvise the execution time for processing of data. An algorithm is implemented namely Scalable and Robust Truth Discovery (SRTD) to detect truth factor in the tweets. This depends on several scores. Here, there are three factors which are used for the efficient truth discovery:Disinformation spread (dispersion of false claims), Data Paucity (lack of evidence from the large dataset), SRTD Algorithm.The above two points can be computed using 'Credibility Score'. With the sum of Attitude Score, Uncertainty Score, Independent Score we will get a Credibility Score, and with that we will get a Reliability Score using SRTD algorithm, if reliability is high then the tweet is true otherwise false. This work is successfully implemented using java.

Key Words:Disinformation spread, Data Paucity, STRD Algorithm

1.INTRODUCTION

Data mining analyzes vast amounts of information to spot market intelligence that corporations will use to unravel problems, mitigate risks, and see new opportunities. The name of the branch of data science was inspired by finding useful information in large databases similar to mountain mining. Both methods search large amounts of data to determine its hidden meaning. Sales and marketing, product development, healthcare, and education are just some of the research and industrial applications that use data mining. If used properly, data mining can help you learn more about your customers, develop effective marketing strategies, increase sales and reduce costs, thus giving you a competitive advantage.

1.1Fundamental Concepts: A variety of tools and techniques are needed to get the promising outcomes from data mining. Some of the most widely used functions are as follows:

- Cleaning and Preparing Data — It refers to the method of changing information into a format appropriate for any analysis and processing, like locating and removing inaccuracies and incomplete data.
- Machine Learning—Refers to programming of computer to allow it to "read" by using statistical probabilities.
- Artificial Intelligence—Planning, listening, reasoning, and problem-solving are among the cognitive tasks performed by these systems, which are close to those performed by humans.
- Clustering—A method of breaking a data set into a set of meaningful groupings to help users understand the structure of their data.
- Classification—In order to correctly predict the target class for each data occurrence, this technique assigns objects in a data set to target groups or categories.
- Regression— Based on a particular data set, a technique for forecasting a range of numeric quantities, such as sales and the stock prices.

2. Literature Survey Review

2.1 Sentiment Analysis for Social Media: A survey:
Harshali P. Patil, et.al has attempted to comprehend the various techniques used in Sentiment Analysis, also known as opinion mining, to define and categorise data on the web ,based on polarity, i.e. positive or negative. The major challenges were the incremental approach, parallel computing for massive data, sarcasm, and grammatically incorrect words. The approach is primarily based on Machine Learning, which allows Sentiment Analysis to be used on a variety of social media sites.

2.2 False News On Social Media: A Data-Driven Survey:
Francesco Pierri, et.al has published a comprehensive report on recent developments in the identification, characterization, and mitigation of false news that spreads through social media. A data-driven methodology is used, with each study's features being classified to characterize false information. The first challenge was separating true from false news articles; the second was the pace and amount at which false news is produced; and the final was social media platforms' restrictions on collecting public data. Undoubtedly, the study claims that this dimension has promising future directions.

2.3 Deep learning for misinformation detection on online social networks: a survey and new perspectives:

Md Rafqul Islam, et.al provided a comprehensive study on automated misinformation identification, including false facts, rumours, spam, fake news, and disinformation. Deep learning is used to automatically process data and generate trends to make decisions not just to extract global features but also to produce better performance, according to a state-of-the-art study on misinformation detection. There are some unresolved problems that limit the idea's execution and future potential.

2.4 Mood-Sensitive Truth Discovery For Reliable Recommendation Systems in Social Sensing:
Jermaine Marshal, et.al proposed a new theoretical paradigm for solving the truth discovery problem that takes mood sensitivity into account. The earlier progress had a flaw in that it didn't look at the mood sensitivity part of the issue. The final result showed that the new model outperformed the baselines in terms of identifying right and mood neutral statements.

2.5 Literature Survey on Sentiment Analysis of Twitter Data using Machine Learning Approaches:
Prashant B. Sawant, et.al proposed a method for pre-processing a stream of tweets from the Twitter microblogging platform and then classifying them based on sentiment. The output of an unsupervised algorithm is investigated. The comparison of the existing system with the proposed system revealed the disadvantages, such as a synonym work vector and results based on assumptions. The disadvantages had a direct effect on algorithm performance. The proposed method is said to solve all of the disadvantages and increase algorithm performance.

2.6 A Literature Review on Twitter Data Analysis:
Hana Anber, et.al looked at a variety of knowledge analysis techniques, including hashtag analysis, Twitter's network topology, event distribution across the network, and impact detection. Because of its simplicity, this model had no significant flaws. Studying the data and its properties, as well as exploring modelling techniques to determine the frequency distribution for each case, will be the subject of future work on this model.

2.7 The analysis of advantages and disadvantages of use of social media in European Union:
Martina Drahoová, et.al conducted research on the three most common social media platforms in the European Union, as well as social media users' perspectives on the benefits and drawbacks of using them. According to all respondents, the most serious disadvantage is Internet addiction. According to 72.2 percent of EU respondents, this is the case. Lack of protection (61.1%), information overload (58.3%), and loss of social contacts are the next most common complaints (47.2 percent). These drawbacks can be mitigated in the future by enhancing security and knowledge.

2.8 Tweet, Truth and Fake News: A Study of BJP's Official Tweeter Handle:
Dr.Amit Sharma, et.al conducted research to better understand the BJP's official tweets' policy, misinformation, media priming, and media analysis. The study's goals are to learn about media priming, truthfulness, and media analysis of the BJP's official tweets. This research can be applied to a variety of social media sites in order to determine the accuracy of the information shared.

2.9 A study on impact of social media over youth of India:
AbhaniDhara K , et.al has investigated the effect of social media on India's youth. The survey results were used to conduct the research. The survey asked questions such as whether social media is useful for education, privacy, information, and other purposes. The key goals were to raise awareness about social media's effect, to clarify the findings about social media's positivity or negativity, to determine the level of social media use, and to determine the most important reason for using social media. The only disadvantage was people's uncivil cooperation in filling out the survey, which may have a subjective effect at times.

2.10 The Impact of Social Media on the Academic Development of School Students:
Tarek A. El-Badawy, et.al has analyzed that Students' academic growth is influenced by social media, according to Questionnaires that were distributed via Facebook and e-mail to see whether social media has an effect on students' academic results. The limited number of people contacted to assess the impact of social media on Egyptian youth is a limitation of this study. After conducting the analysis, it was determined that social media has little impact on school students' academic performance because, despite spending hours on social media, they still manage to study and earn good grades.

3. IDENTIFICATION OF PROBLEM

Existing System
Truth discovery is a crucial task in social media anticipation, in which the intention is to identify credible origin and the factual statements from significant noisy, unfiltered, or evena few contradictory social media information. To address the authenticity, the experts in machine learning and data mining have proposed some principled approaches.

Disadvantages of Existing System:
1. Present truth discovery approaches do not specifically explain the "false information spread" problem, in which a large number of sites spread misleading information on social media platforms.
2. Several modern reality discovery algorithms depend extra closely on correct evaluation of source reliability, which often calls for a densely packed dataset.

Proposed System :
It has been substantially mentioned a Scalable and Robust Truth Discovery additionally referred to as SRTD system in this paper to tackle the dissemination of disinformation, data sparsity, and problems for scalability of different social media applications. To counter the issue of misinformation distribution, the SRTD scheme specifically models different source behavior such as duplicating, self correction, and the spam data that is available on the internet. To cope with statistics deficiency, the SRTD scheme employs a brand new method that measures declaration truthfulness primarily based

totally on a reliability evaluation of the claims concerning content material in addition to the historic contributions of beginning that result in the claim. The findings are compelling due to the fact they offer a sturdy and flexible method to clear up the fact discovery problem on social media platforms with noisy and insufficient statistics.

4. IMPLEMENTATION

4.1Scalable and Robust Truth Discovery (SRTD)

In today's generation where social media provides a new paradigm in which people function as pervasive, inexpensive, and scalable sensors, spontaneously reporting their findings of the physical environment. The popularity of compact data collection devices, as well as the vast data distribution opportunities provided by social media, are driving this paradigm. Truth discovery is a crucial task in the social media sense, to identify credible origin and legitimate statements from vast noisy, unprocessed, and even contradictory data on social media. Truth discovery issue remains at the core challenge of social media applications.

4.2 JACCARD ALGORITHM FOR INDEPENDENT SCORE

Jaccard similarity index compares components from two different sets to identify which ones are common and which are unique. It is a scale that ranges from 0% to 100% for determining how close two sets of data are. The closer the two populations are, the greater the ratio. It is easy to use, but it is highly sensitive to small samples and therefore can produce wrong output, especially with very small samples or databases with missing data.

Formula : $\mathbf{J(X,Y) = |X \cap Y| / |X \cup Y|}$

4.3 STANFORD CORE NLP FOR SENTIMENT ANALYSIS

Users may use CoreNLP to generate annotations for text in linguistics, such as token, sentence borders, person's name, quantitative and time values, Parse dependencies and constituencies, coreference, sentiment and connections. Arabic, Chinese, English, French, German, and Spanish are currently supported by CoreNLP. Stanford CoreNLP is a Java natural language analysis library that provides analytical NLP, machine learning NLP, and rule based NLP tools for major computing problems and can be incorporated into various applications that use human speech technology.

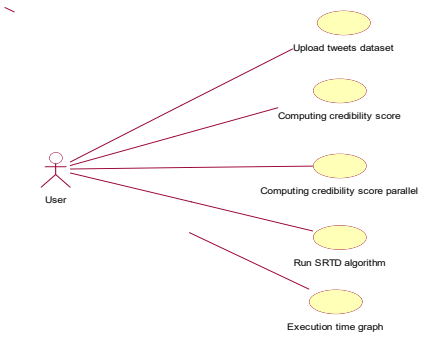

Fig -1: Use Case Diagram

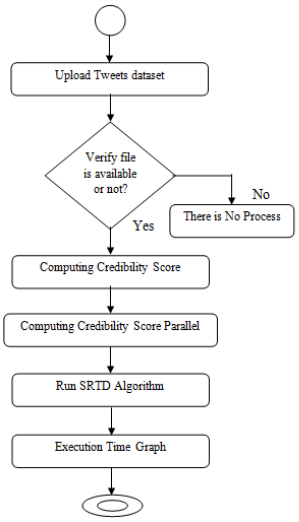

Fig -2: Activity Diagram

Table -1: Test Cases

Id	Name	Description	Test Case Steps			Status	Priority
			Step	Expected	Actual		
1	Upload Dataset	Verify dataset available or not	If not available	No process	Dataset loaded	High	High
2	Computing Credibility Score	Verify dataset loaded or not	If not loaded	Score cannot be calculated	Calculated score (Normal)	High	High
3	Computing Credibility Score Parallel	Verify dataset loaded or not	If not loaded	Score cannot be calculated	Calculated score (Parallel)	High	High
4	Run SRTD Algorithm	Verify score calculated or not	If not calculated	Cannot run the algorithm	Tweet reliability score is seen	High	High
5	Execution Time Graph	Verify dataset available or not	If not available	Graph can't be found	execution time in graph can be seen	High	High

5. COMPARISION

- There is no major drawback to this system as it has various future directions and is not only limited to this platform.
- Due to the simplicity of this system, it can be used for various platforms and there is always a room for improvements that leads to upgradation of the system

6. FUTURE SCOPE

- The SRTD is a simple system which is proposed to overcome the drawbacks of truth sensing in big data concepts. Earlier systems processed one unit of data at one time which impacted the efficiency of system in a poor manner leading to more load and more time.
- This system allows input of massive data as it has parallel method of calculation, which concurrently processes several units of data at once. The parallel method is one the biggest advantage which enables the boundless characteristic [in terms of platform] of SRTD.
- This approach can be expanded to diverse social media systems like Instagram, Twitter, Telegram, Facebook, WhatsApp, LinkedIN etc.

- The system can be successively improved by using API to extract real-time data and then process that data for truth discovery.

7. CONCLUSION

In this we suggested an algorithm known as Scalable Robust Truth Discovery to find a solution for the data truthfulness issue in popular various social media platforms. To effectively solve the disinformation dissemination and the data insufficiency threats in the discovery of truth crisis, we specifically considered source authenticity, report integrity, and a source's historical activities in our approach. To solve the problem's scalability issue, we have developed a distributed. The real world data was obtained from Twitter and it was used to validate the SRTD algorithm. The resultsshow that our solution performs better in terms of authenticity and computational efficiency. The results are compelling because they provide a powerful and flexible way to solve the problem of finding the truth in noisy and inadequate data that is provided on various social media platforms.

ACKNOWLEDGEMENT

We would like to express our deepest gratitude to our guide, Ms. Sonam Sharma for her valuable guidance, consistent encouragement, personal caring, timely help andproviding me with an excellent atmosphere for doing research. All through the work,in spite of her busy schedule, she has extended cheerful and cordial support to us forcompleting this research work.

REFERENCES

[1] Patil, Harshali&Atique, Mohammad. (2015). Sentiment Analysis for Social Media: A Survey. 1-4. 10.1109/ICISSEC.2015.7371033.

[2] Pierri, Francesco &Ceri, Stefano. (2019). False News On Social Media: A Data-Driven Survey. ACM SIGMOD Record. 48. 18-32. 10.1145/3377330.3377334.

[3] Islam, Md Rafiqul & Liu, Shaowu& Wang, Xianzhi & Xu, Guandong. (2020). Deep learning for misinformation detection on online social networks: a survey and new perspectives. Social Network Analysis and Mining. 10. 10.1007/s13278-020-00696-x.

[4] Marshall, Jermaine & Wang, Dong. (2016). Mood-Sensitive Truth Discovery For Reliable Recommendation Systems in Social Sensing. 167-174. 10.1145/2959100.2959147.

[5] B. Sawant, P., 2017. *Literature Survey on Sentiment Analysis of Twitter Data using Machine Learning Approaches*. [online] .

[6] Anber, Hana &Akram Salah, & Ahmed, Abd El-Aziz. (2016). A Literature Review on Twitter Data Analysis. International Journal of Computer and Electrical Engineering. 8. 241-249. 10.17706/IJCEE.2016.8.3.241-249.

[7] Drahošová, Martina & Balco, Peter. (2017). The analysis of advantages and disadvantages of use of social media in European Union. Procedia Computer Science. 109. 1005-1009. 10.1016/j.procs.2017.05.446.

[8] Sharma, Amit & Goyal, Aayushi. (2018). Tweet, Truth and Fake News: A Study of BJP's Official Tweeter Handle. Journal of Content, Community and Communication. 4. 22-28. 10.31620/JCCC.12.18/05.

[9] AbhaniDhara K, "A study on impact of social media over youth of india", International Journal of Engineering Development and Research (IJEDR), ISSN:2321-9939, Vol.7, Issue 2, pp.24-33, April 2019 .

[10] Hashem, Yasmin. (2015). The Impact of Social Media on the Academic Development of School Students. International Journal of Business Administration. 6. 46. 10.5430/ijba.v6n1p46.