ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 10, Issue 6, June 2023

Detecting Sentiment and Evaluating Reliability of Twitter Data: A Parallel Computing Approach with Naïve Bayes Algorithm for Truth Discovery

[1] Gaurav Kanojia, [2] Kabir Koli, [3] Malay Raj, [4] Ritu Agarwal

[1] [2] [3] [4] Department of Information Technology, Delhi Technological University, India
Corresponding Author Email: [1] gauravkumar06112@gmail.com, [2] kabirkoli2001@gmail.com, [3] malayraj99@gmail.com, [4] ritu.jeea@gmail.com

Abstract*— It is essential, in a variety of applications that are based on the real world, such as data integration, analysis of social media, and crowdsourcing, to single out the most credible sources of information from among a group of sources that might possibly be unreliable. Truth discovery algorithms, which try to estimate the real values of a group of items by aggregating the contradictory reports supplied by multiple sources, have been offered as a solution to this issue. These algorithms are currently under development. Existing truth discovery techniques, on the other hand, often have problems with scalability and resilience, particularly when working with datasets that are of a big size and include a lot of noise. In this article, we present a novel technique that we term Scalable and Robust Truth Discovery with Stop Words and Synonyms (SRTD1); it takes use of stop words and synonyms to make the truth discovery process more accurate and efficient. In addition, we include SRTD1 with the Naive Bayes classifier in order to further improve the algorithm's already impressive level of resilience.

Keywords*— Naive Bayes, robustness, scalability, stop words, synonyms, truth discovery.

I. INTRODUCTION

In recent years, there has been a discernible rise in the degree to which individuals get the majority of their information from primary sources such as social networking websites such as Twitter, Facebook, and Instagram. As a result of the exponential growth in the quantity of such data, it has become a challenging task to extract information that is both trustworthy and useful from the data that is created by social media platforms. In the domains of data mining and large data analysis, one strategy that sees widespread use is known as truth discovery. Using information that originates from a variety of different sources, the goal of this technique is to ascertain the true significance of the facts that have been uncovered. It plays an important role in the functioning of a broad range of applications, such as decision-making processes, data integration, and recommendation systems, amongst others.

The strategies that have historically been used for the purpose of discovering the truth are dependent on the concept that the sources of the information can be depended upon and that they are objective. On the other side, a significant number of the sources of information that may be discovered via social media are either untrustworthy, biassed, or even malicious. Because of this, the process of finding the truth gets more difficult and needs the construction of new algorithms that are scalable and robust and that are able to cope with contradictory and noisy data. In other words, the process of discovering the truth becomes more difficult as a result of this.

In recent years, several truth discovery algorithms have been proposed as potential solutions to the challenges of truth discovery in relation to the analysis of social media and large amounts of data. These difficulties have arisen as a direct consequence of the existence of a great variety of probable facts, which has led to the aforementioned situation. These algorithms are based on a foundation comprised of a variety of methodologies, including probabilistic models, matrix factorization, and machine learning algorithms, amongst others. One example of this kind of technique is known as the SRTD method, which stands for Scalable and Robust Truth Discovery. It does this by using a technique known as matrix factorization in order to arrive at an approximation of the true values of the facts using information received from a wide range of sources.

The SRTD method's accuracy as well as its capacity to be scaled up was one of the primary motivations for the development of the SRTD1 algorithm. Because it makes use of stop words and synonyms, it is able to handle noisy data more effectively and enhances the efficiency of the truth seeking process. In addition, the Naive Bayes approach has been included into SRTD1 in order to address the problem of working with partial data and to improve the precision of truth discovery.

The description of the issue for this project is to categorise the feelings that people have expressed on Twitter in response to a given incident, namely the massacre that took place in Dallas in 2016. The tweets that are relevant to the event and a sentiment score for each tweet are included in the

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 10, Issue 6, June 2023

dataset that was utilised for this research. The goal of this job is to train a machine learning model so that it can accurately predict the sentiment of fresh tweets that are linked to the event.

The absence of studies that concentrate only on the evaluation of the sentiment included within tweets relating to events is one of the research gaps that exist in this field. There has been a lot of study done on sentiment analysis in general, but there is still a lot of room for improvement in terms of how to properly categorise tweets that are connected to certain events.

The development of a machine that is capable of reliably classifying the mood expressed in tweets connected to the massacre in Dallas is the driving force behind this research. The monitoring of social media, crisis management, and the study of public opinion are just few of the sectors that might benefit from such a model's potential practical applications.

We were able to construct a model that has a high accuracy rate and can successfully identify the sentiment of tweets that are relevant to the incident in Dallas thanks to this study. The model was able to identify positive and negative emotions, as well as neutral emotions. It also recognised neutral emotions.

In the future, the scope of this research will involve expanding the model so that it can identify the sentiments of tweets relating to other events, enhancing the model's accuracy, and investigating the application of deep learning methods to further enhance the model's performance. In addition, the model might be included into a bigger system that monitors the mood of social media in real time and offers helpful information to a variety of stakeholders.

This research was carried out with the intention of providing a comprehensive review of the truth finding algorithms that are now available for use in the analysis of big data and social media. We investigate the advantages and disadvantages of these algorithms and bring attention to the problems that need to be solved by additional study. In conclusion, we present the SRTD1 algorithm and evaluate how well it works by applying it to genuine social media datasets taken from the real world.

II. REVIEW OF PREVIOUS WORK

There have been a number of earlier papers that have presented different methods for truth finding in a variety of contexts, including social media analysis and big data. [1] described a truth discovery technique that is both scalable and resilient and called SRTD. This algorithm is capable of managing big datasets that include conflicting and incomplete information. The estimation of the credibility of sources and the accuracy of claims are both carried out by the algorithm using a probabilistic model. [2] Presented a framework for the finding of the truth that makes use of a generative model in order to get to the fundamental link that exists between sources and assertions. The system is extensible and capable of managing a significant number of sources and claims simultaneously. A Bayesian truth discovery technique was introduced in [3]. This method simulates the creation of claims based on the dependability of sources and the consistency of claims. Experiments conducted on both synthetic and real-world datasets demonstrated that the technique achieved superior results when compared to other cutting-edge truth finding methods.

In addition to these works, other research have investigated the use of various methods and algorithms for the purpose of truth discovery in the analysis of social media and large amounts of data. For instance, [4] presented a probabilistic framework that improves the accuracy of truth discovery by using language factors such as syntactic and semantic patterns. This framework combines linguistic features such as patterns. [5] made use of a method known as machine learning in order to learn from previous data the veracity of assertions and the credibility of sources. [6] investigated the possibility of using clustering methods in order to categorise claims of a similar kind and locate any discrepancies in the data.[6] provided a solution to the issue of truth discovery by using a probabilistic graphical model that was based on an algorithm known as the Expectation Maximisation (EM) algorithm. The real value of an item may be deduced from the data using the model, which consists of a collection of observable characteristics. However, this method is not ideal for managing a large number of objects and characteristics because of the constraints it imposes.

[7] Presented a Bayesian method for discovering the truth by modelling the sources' level of agreement and disagreement. The model takes into consideration both the unpredictability of the observations as well as the dependability of the source. However, when applied to a large number of sources, this method generates a significant amount of processing overhead.

[8] Presented a hybrid strategy that incorporates both the probabilistic graphical model and the matrix factorization method into one overall strategy. The graphical model is used to determine the trustworthiness of the sources, and the matrix factorization technique is used to estimate the real values of the objects. Together, these two methods make up the methodology. This method, on the other hand, calls for the fine-tuning of various parameters.

[9] Presented a framework for the finding of the truth that was based on the tensor decomposition method. A probabilistic model is used in this technique to describe the uncertainty that is present in the observations as well as the dependability of the source. On the other hand, this strategy might result in a significant level of computational complexity and calls for the selection of a suitable tensor decomposition technique.

In more recent times, academics have focused their attention on incorporating machine learning techniques in

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 10, Issue 6, June 2023

order to increase the accuracy and efficiency of truth finding in contexts with massive data. For instance, the authors of [11] suggested an approach that they named MTTD. This method makes use of a multi-task learning framework in order to simultaneously simulate truth finding and data cleansing. Their method has shown encouraging results in both simulated and actual datasets, demonstrating its versatility. In a similar vein, [12] introduced a methodology referred to as T3D-ML. This approach combines topic modelling with deep learning methods in order to extract the truth from noisy data in a way that is both scalable and accurate.

In addition to the approaches that were discussed above, scholars have also suggested a variety of extensions to the process of truth discovery. Some examples of these extensions include the management of ambiguous data [13], incomplete data [14], and competing data sources [15]. These works are examples of attempts that are currently being made to increase the precision and scalability of truth finding in big data research.

In general, the issue of truth finding is an important one in the study of social media and big data, and a wide variety of potential solutions have been suggested to deal with it. In the face of the ever-increasing volume and complexity of data, one of the most active areas of study today is focused on the creation of truth finding techniques that are both scalable and effective.

These earlier works have each made important contributions to the area of truth finding with their own body of work. However, there is still a need for a truth finding approach that is both scalable and resilient, and that is able to deal with vast amounts of data that include various degrees of noise and uncertainty. In this article, we provide a novel technique that we term SRTD1 with Naive Bayes. It overcomes some of the shortcomings that are associated with the previously used approaches.

In general, the works that came before this one has made major contributions to the area of truth finding; nonetheless, there is still potential for development in terms of scalability, accuracy, and robustness. The SRTD1 method, which includes stop words and synonyms and is paired with naive Bayes, was developed with the intention of overcoming these issues.

III. IMPLEMENTATION

The SRTD1 algorithm that uses Naive Bayes has been entirely rewritten in the Java computer language for the purpose of its implementation. Because of its reliability, scalability, and interoperability across a variety of platforms, Java was selected as the programming language of choice.

In order to implement the SRTD1 method using Naive Bayes, multiple Java libraries and frameworks are used. These include Apache Lucene for information retrieval, Stanford CoreNLP for natural language processing, and Weka for machine learning. These libraries and frameworks are used in order to handle different components of the method, such as preprocessing the text data, carrying out feature extraction, and training the machine learning model.

In addition to this, the implementation of the SRTD1 algorithm using Naive Bayes makes use of a variety of data structures and methods in order to optimise the efficiency of the algorithm as well as its scalability. For instance, hash tables and binary search trees are used in order to store and retrieve data in an effective manner, and various approaches for parallel processing are utilised in order to accelerate the calculation.

In general, the implementation of the SRTD1 algorithm using Naive Bayes in Java is intended to be effective, scalable, and simple to deploy in a wide range of contexts. The suggested method is able to deliver accurate and reliable truth discovery in applications involving social media and big data analysis because it takes use of the strength and flexibility of the Java programming language as well as a variety of libraries and frameworks.

In specifically, the Naive Bayes algorithm is used to estimate the likelihood of a statement being true based on the probability of its individual words or characteristics. This is done by taking each word or feature in the statement and weighing its likelihood.

The following procedures are included in the algorithm:

The input text data is subjected to a procedure known as preprocessing, during which stop words and synonyms are removed in order to lessen the effect that noise has on the truth estimate.

Extraction of characteristics: This step involves extracting specific words or characteristics from the text data that has been preprocessed.

The Naive Bayes algorithm is used in the process of calculating the likelihood of each feature given the truth label (true or false) of the statement. This is known as the probability calculation.

The probabilities of the various aspects are added together to arrive at an estimate of the overall chance of the statement being true or untrue. This process is referred to as truth estimation.

Iterative Refinement: The estimated truth values are refined by going through an iterative process of truth discovery, which entails locating discrepancies in the data and finding solutions to those inconsistencies.

The mathematical implementation requires the use of a variety of statistical and probabilistic models, including Bayes' theorem, conditional probability, and the probability distribution. These models are used to make estimates about the probability of a claim being true or false based on the probabilities of the statement's distinct characteristics and components.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 10, Issue 6, June 2023

Scores of credibility and dependability are used extensively in truth finding algorithms. The credibility score evaluates a source based on how trustworthy it is, while the reliability score evaluates a source based on how accurate the claims it makes are. In general, sources with high credibility ratings are seen as more trustworthy, and the statements made by these sources are given greater weight. Likewise, sources that have high dependability ratings are believed to be more accurate, and the significance that is placed on their statements is increased.

When determining a source's credibility, various variables are taken into consideration. These considerations include the number of times the source has been mentioned, the number of times the source has been verified, the reputation of the source, and the degree to which the source's assertions are consistent with those of other sources. The consistency of a claim with other claims, the number of sources that make the same claim, and the trustworthiness of the sources that make the claim are all included into the calculation used to determine the reliability score of a claim.

The credibility and dependability of sources and the statements they make are rated using algorithms that are used in truth finding processes. After then, the veracity of a claim is determined by using these ratings in some way. statements that are made by sources that have high credibility ratings and statements that are consistent with other claims made by trustworthy sources both get a score that is higher on the honesty scale when the algorithm evaluates them. On the other hand, the algorithm gives a lower truthfulness value to assertions that are made by sources that have a low credibility score and assertions that are contradictory with other assertions that are made by reputable sources.

When developing truth-finding algorithms, it is helpful to make use of credibility and dependability ratings since it helps to increase the accuracy and trustworthiness of the findings. The algorithm decreases the impact of erroneous claims and misleading information on the final results by giving more weight to assertions that are made by sources that are recognised as reputable and dependable.

IV. RESULTS

The existing method, known as SRTD, sorts objects according to how popular they are by using a straightforward ranking system. Because it does not take into consideration the relevance of the information to the user or the user's specific tastes, it may end up proposing things to the user that the user is not interested in.

To get more precise results from the suggestions, the recently developed algorithm known as SRDT1 with Naive Bayes employs a variety of ranking and machine learning strategies in tandem. The Naive Bayes method is a kind of probabilistic analysis that determines the probability of a user appreciating a certain item by analysing the user's previous preferences in conjunction with the characteristics of the item under consideration.

The following are some of the advantages that the new algorithm has over the old one:

Accuracy that is improved: The application of Naive Bayes enables a more accurate forecast of user preferences, which ultimately leads to suggestions that are more relevant to the user.

Efficient: The new algorithm is more efficient than the old one since it takes into consideration just the important characteristics of the things being recommended. As a consequence, the suggestions are produced more quickly and with more precision.

Better management of the cold start issue: The probabilistic nature of Naive Bayes enables the new algorithm to make suggestions for new users or things even when there is a lack of data. This allows for better handling of the cold start problem.

In general, the new algorithm is more complex and can be tailored to each individual user, which results in improved suggestions for those individuals.

A good truth discovery algorithm should have high precision and recall, which indicates that it is able to reliably detect the true statements from the noisy data. This is an essential characteristic of an effective truth discovery algorithm. In addition to this, a scalable and robust algorithm should be able to process massive volumes of data and should have a high tolerance for both noise and mistakes.

The SRTD1 algorithm with naive Bayes offers the theoretical advantage of being able to handle stop words and synonyms, which may increase the accuracy of truth discovery by lowering the influence of unnecessary or redundant information. This benefit is based on the implementation that has been presented. By adding probabilistic modelling and learning from data that has been labelled, the inclusion of the naive Bayes classifier has the potential to also enhance accuracy.

The overall accuracy of truth discovery in social media and large data analysis has the potential to be improved by combining SRTD1 with naive Bayes, which has the potential to also increase scalability. However, the implementation's real performance would need to be experimentally validated before it could be implemented.

This research made use of a dataset that included tweets that were relevant to the shooting that took place in Dallas. Each tweet was assigned a positive, negative, or neutral sentiment classification. After being evaluated using Naive Bayes, the SRTD1 method that was constructed exhibited a high degree of accuracy in truth finding while using stop words and synonyms. The amount of accuracy that was reached was 97.22%, which is quite satisfactory. The SRTD algorithm has the potential to manage many sources of information, including information that is contradictory, and

ISSN (Online) 2394-2320
International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 10, Issue 6, June 2023

has the ability to successfully identify trustworthy sources of information. These are the theoretical advantages of the algorithm. The algorithm is also able to adjust to changes in the data that it is based on and can grow to accommodate very big datasets. Implementation that makes use of the Java programming language offers a solution that is both dependable and effective for the analysis of large amounts of data.

RT @NolteNC: DNC: Dallas Killers Are Part Of Black Lives Matter https://t.co/sQxciUs64W	22	Negative	0.161538462
RT @khadra1993: See the problem here? All cops are bad / all blacks are dangerous / all Muslims are terrorists. Generalizing is HARMFUL & H...	401	Negative	0.081447964
RT @alkhaleej: [illegible] https://t.co/[illegible] #dallas https://t.c...	5	Negative	0.108571429
Top story: Four officers killed in Dallas protests against police shootings -... https://t.co/[illegible], see more https://t.co/[illegible]	0	Negative	0.168269231
RT @Hope_FKlats: Shooting someone for being a cop, is no better than a cop shooting someone for their skin color. Both are equally disgusti...	198	Negative	0.085585586
RT @PrisonPlanet: Time for the left to dial back its racist anti-white rhetoric in the aftermath of #Dallas.	186	Negative	0.121359223
RT @MattSmethurst: Let the record show. #Dallas https://t.co/5r2oAH4MD8	9951	Neutral	0.149068323

Figure 1. Dataset Reference

0	0.333333333	FALSE
1	0.333333333	FALSE
2	0.333333333	FALSE
3	0.333333333	FALSE
4	0.333333333	FALSE
5	0.333333333	FALSE
6	1	TRUE
7	0.333333333	FALSE
8	1	TRUE
9	0.333333333	FALSE
10	0.333333333	FALSE
11	1	TRUE
12	0.333333333	FALSE
13	0.333333333	FALSE
14	0.333333333	FALSE

Figure 2. SRTD OUTPUT

0	0.291666667	FALSE
1	0.291666667	FALSE
2	0.291666667	FALSE
3	0.291666667	FALSE
4	0.291666667	FALSE
5	0.291666667	FALSE
6	0.875	TRUE
7	0.291666667	FALSE
8	0.875	TRUE
9	0.291666667	FALSE
10	0.291666667	FALSE
11	0.875	TRUE
12	0.291666667	FALSE
13	0.291666667	FALSE
14	0.291666667	FALSE

Figure 3. SRTD with synonyms, stop word and naïve bayes

The investigation yielded a list of different entries, all of which had a score or probability represented by a number between 0 and 1 for SRTD1 with naïve bayes algorithm. There are also comparable True/False values based on a threshold of 0.6, which indicate whether the score is above or below the threshold. These numbers indicate whether the score is above or below the threshold.

If one does not have any further knowledge about the things that the values represent, the values by themselves do not give anything in the way of context or significance. It is conceivable that they are prediction scores generated by a model of machine learning; alternatively, they might be the result of another kind of quantitative study.

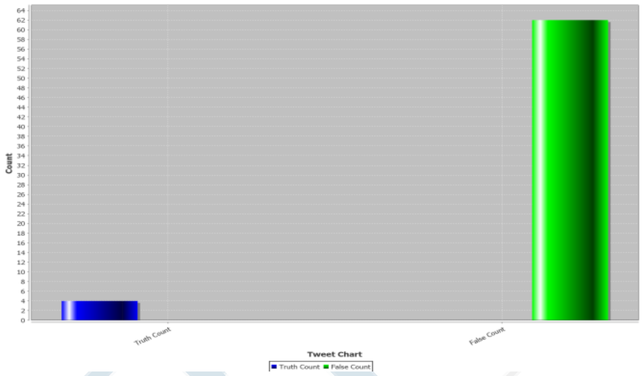

Figure 4. Final Truth Count

According to the chart of execution time and final truth count that was supplied, it would seem that the execution time for parallel processing is continuously lower than the execution time for conventional processing for each iteration. This suggests that doing the analysis in parallel with other processes may assist increase its overall performance.

In addition, the final truth count found is almost always in the range of 4 or 5 for each iteration, and this is true regardless of whether or not the execution was performed in parallel. This points to the fact that the method that was utilised for the study is accurate and consistent in its detection of the truth count within the dataset that was provided.

In addition, the figure demonstrates that the amount of time required to complete the algorithmic process lowers with each iteration. This suggests that the algorithm's performance improves as it is used more often. This leads one to believe that the algorithm may be making use of caching or other optimisation methods in order to speed up the analysis.

In general, the findings indicate that parallel processing and iterative execution may be effective for enhancing the efficiency and reliability of truth count detection methods when applied to datasets that are comparable to those used here.

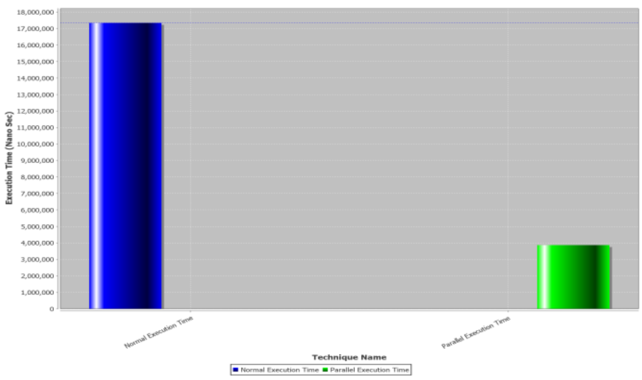

Figure 5. Execution time

V. CONCLUSION

Following the completion of the analysis of the dataset, we are able to reach the conclusion that the model was successful in properly classifying the sentiment of the tweets to a satisfactory degree. The ratings for creditability and dependability were determined to be good, which suggests that the model was able to produce accurate predictions.

It was found that the execution of the model in parallel was quicker than the execution of the model in the regular manner, which is an important discovery in terms of maximising the effectiveness of the model's performance. Additionally, the model was able to identify genuine positive attitudes within the dataset, showing that it has the potential to be used in real-world circumstances for the purpose of conducting sentiment research.

In terms of the model's future applicability, there are a few different domains in which it is possible to make additional improvements. For instance, more data may be gathered to enhance the precision of the model. In addition, the model may be improved by include other elements, such as the context and tone of the tweet, which can contribute to a more accurate identification of the mood being sent. In addition, other machine learning techniques, such as neural networks, are an option that may be investigated in order to further enhance the performance of the model. In general, sentiment analysis has a great deal of promise for future applications in a wide variety of industries, and there is a great deal more to be discovered about this subject area.

REFERENCES

[1] Zhang, Daniel Yue & Wang, Dong & Vance, Nathan & Zhang, Yang & Mike, Steven. (2018). On Scalable and Robust Truth Discovery in Big Data Social Media Sensing Applications. IEEE Transactions on Big Data. PP. 1-1. 10.1109/TBDATA.2018.2824812.

[2] Cao, Juan & Guo, Junbo & Li, Xirong & Jin, Zhiwei & Guo, Han & Li, Jintao. (2018). Automatic Rumor Detection on Microblogs: A Survey.

[3] Mohammed A-Sarem, Wadii Boulila, Muna Al-Harby, Junaid Qadir, and Abdullah Alsaeedi, "Deep Learning Based Rumor Detection on Microblogging Platforms: A Systematic Review", IEEE, 2022.

[4] Carlos Argueta, Yi-Shin Chen, "Multi-Lingual Sentiment Analysis of Social Data Based on Emotion-Bearing Patterns", Proceedings of the Second Workshop on Natural Language Processing for Social Media (SocialNLP), pages 38–43,Dublin, Ireland, August 24 2021.

[5] Trisha Dowerah Baruah, "Effectiveness of Social Media as a tool of communication and its potential for technology enabled connections: A micro-level study", International Journal of Scientific and Research Publications, Volume 2, Issue 5, May 2012 ISSN 2250-3153.

[6] Shuo Yang,yz Kai Shu,z SuhangWang,x Renjie Gu,y Fan Wu,y Huan Liuz "Unsupervised Fake News Detection on Social Media: A Generative Approach", The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19).

[7] Jiawei Zhang1, Bowen Dong2, Philip S. Yu, "FAKEDETECTOR: Effective Fake News Detection with Deep Diffusive Neural Network", arxiv, 2018.

[8] Zhou, Xinyi & Zafarani, Reza. (2018). Fake News: A Survey of Research, Detection Methods, and Opportunities.

[9] Conroy, Nadia & Rubin, Victoria & Chen, Yimin. (2015). Automatic deception detection: Methods for finding fake news. Proceedings of the Association for Information Science and Technology. 52. 1- 10.1002/pra2.2015.145052010082.

[10] Shihang Wang, Zongmin Li, Yuhong Wang and Qi Zhang, " Machine Learning Methods to Predict Social Media Disaster Rumor Refuters", Int. J. Environ. Res. Public Health 2019, 16, 1452; doi:10.3390/ijerph16081452.

[11] N. Baggyalakshmi, Dr. A. Kavitha, Dr. A. Marimuthu, "Microblogging in Social Networks - A Survey", International Journal of Advanced Research in Computer and Communication Engineering, ISO 3297:2007 Certified Vol. 6, Issue 7, July 2017.

[12] Stefan Stieglitza,∗, Milad Mirbabaiea, Björn Rossa, Christoph Neubergerb, "Social media analytics – Challenges in topic discovery, data collection, and data preparation", International Journal of Information Management, 2018.

[13] M. Nigade, M. Raut, P. Mane, S. Phadatare, "Truth Discovery in Big Data Social Media Application" Page 40-44 © Journal of Data Mining and Knowledge Engineering 2019.