Addressing misinformation spread and uncovering spammersusing social honeypot in truth discovery on social media.

Dhanave Pallavi
Department of Computer Engineering

SPVP's S.B. Patil College of Engineering
Indapur, Pune

Mira Bhavar
Department of Computer Engineering

SPVP's S.B. Patil College of Engineering
Indapur, Pune

Tulsa Chavan
Department of Computer Engineering

SPVP's S.B. Patil College of Engineering
Indapur, Pune

Aruna Salunke
Department of Computer Engineering

SPVP's S.B. Patil College of Engineering
Indapur, Pune

***Abstract*—The sources of data from online social media may be consist of some data which are noisy and sparse. While handling of big data related social sensing media application their challenges like misinformation spread are data sparsity and fake news. The system is going to use of advanced algorithms to discover the dynamic truth information and frequently used information. The addressing misinformation spread in big data (e.g. whatsaap, twitter, instagram) is difficult task in the present days. There will be one more challenge is data sparsing, where majority of sources contributes only small number of claims. The existing solution are not enough for large scale social sensing events, since existing algorithms have centralized in nature. We are going to use of the SRTD scheme for identifying both source realibility as well as crediability of the claims. In the honeypots system we can For implementation of the system we use the social honeypot system to find out the spammers and legitimates which can upload the tweets on social media. By using the honeypot system we analyze the different tweets which are true or false. In truth discovery significant amount of attentions has recieved in recent years and we will be developing various models to address truth information.**

Keywords*— *Big Data, Truth Discovery, Scalable, Data Sparsity, Realibility.

I. Introduction

We find that the deployed social honeypots identify social spammers with low false positive rates and that the harvested spam data contains signals that are strongly correlated with observable profile features like content, friend information, etc. This paper presents new truth discovery approach on social media. Lagre number of news releasing on social media application from that all are not true it may be false or fake. In big data social media like facebook, whatsaap, twitter large amount of data which may difficult to discover. There is big challenge in truth discovery i.e. misinformation spreading [1], There may challenge in like in social media application correctness of reported observation and reliability of data sources[5]. Furthermore, unlike claims generating by human which add further complexity to the truth discovery.

Consider the example in which the information of the any famous actor is spreading which is not true, which may be known as "misinformation spread". Also example of social media sensing include real world awareness in intelligent transport system application[5]. Principle of solution from data mining, data analytics and network sensing communities which addressing truthness of problem[1][5]. The main two challenges like "data sparsity" and "misinformation spread" means spreading false information on social media addressing by "on robust truth discovery in sparse social media sensing[1][3]. In such case the finding correctness of claims is little bit difficult.

Recent efforts have been made to solve dynamic truth discovery of problem like noisy and incomplete data, where social media sensing data is sparse in nature [4]. These solutions include new constraint aware dynamic truth discovery schema (CA-DTD) Markov model [1] [4]. This schema recently applied on real world data sets. Note that, a trivial way of accomplishing the truth discovery task is by "believing" only those observation that are reported by a sufficient number of sources[2]. A significant challenge in social media sensing application lies in ascertaining the correctness of collected data. In previous work optimal solution for the truth discovery is made by using maximum likelihood estimation [2] [1]. This observation of current literature of truth discovery is given as above.

A. Objectives

1. To reduce the spreading of misinformation on social media sensing application.
2. To detect the dynamic truth information.
3. To identify truthful claims among widely spread misinformation.
4. To solve the truth discovery problem in big data social media sensing applications.

B. Motivation

The first one is "misinformation spread" where a significant number of sources are contributing to false claims, making the identification of truthful claims difficult. For example, on Twitter, rumors, scams, and influence bots are common examples of sources colluding, either intentionally or unintentionally, to spread misinformation and obscure the truth. The second challenge is "data sparsity" or the "long-tail phenomenon" where a majority of sources only contribute a small number of claims, providing insufficient evidence to determine those sources' trustworthiness.

II .LITERATURE SURVEY

The multimedia social event summarization framework which automatically generates holistic visualized summary from the microblogs of various media types was presented by Jingwen Bian, Yang Yang, Hanwang Zhang, Tat-Seng Chua; they developed three major stages to accomplish the summarization. First, they devised an effective approach for eliminating noisy images from raw collection. Then a novel Cross-Media-LDA (CMLDA) model was proposed. Finally they generated the multimedia summary for social events.
Danial (yue) Zhang, Rungana Han, Dong Wang, Chao Huang, find two fundamental challenges in truth discovery problem in social media sensing. They develop a novel Robust Truth Discovery (RTD) Scheme that explicitly considers both the fine-grained source attitude and source's historical contributions. The RTD addressed the misinformation spread and data sparsity. They introduce the concept of Contribution Score (CS) of sources to address the data sparsity. Two large scale real world data sets were used to evaluating the performance of scheme.
Danial (yue) Zhang, Dong Wang, Yang Zhang, introduced the physical constraint awareness, which was Hard constraints and Soft constraints.They present the Constraint Aware Dynamic Truth Discovery (CA-DTD) scheme which consist two key components: Constraint-Aware Hidden Markov Model (CA-HMM) and Complimentary Source Incorporation (CSI). The system result was important since they lay out solid analytical foundation to address the dynamic truth discovery. The overview of CA-DTD scheme in this survey is as follows:

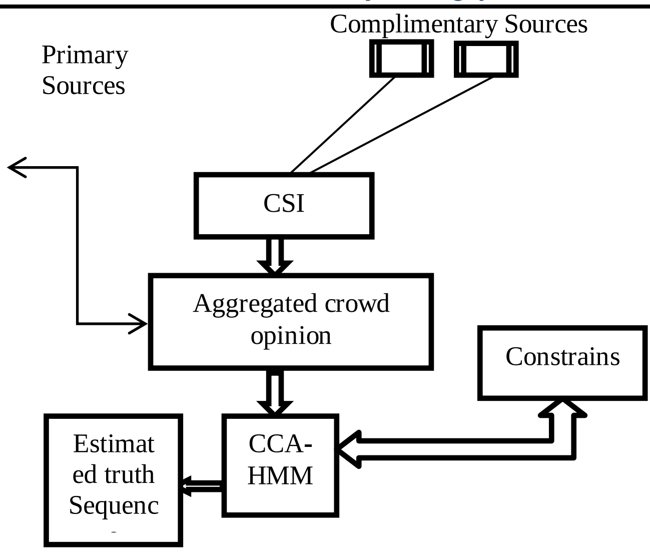


Fig: Overview of CA-DTD Scheme

Dong Wang, Lance Kaplan, Hieu Le, Tarek Abdelzaher, introduced the problem of finding the maimum likelihood estimates of parameters in statistic model, where data is" incomplete". To solve this problem they discovered the Expectation-Maximization (EM) Algorithm. The EM approach can determine the correctness of Reported observations; optimal solution is obtained by solving it. The solution is directly lead to an analyatically founded quantifiacation of the correctness of measurements and realiability of participants.
Chao Huang, Dong Wang and Nitesh V.Chawla, discovered the challenge of finding the reliability of sources without without prior knowledge lies in social sensing application.they founded two main Schemes which was Uncertainty-Aware Truth Discovery (UTD) and Scalable Uncertainty-Aware Truth Discovery (SUTD).The SUTD Scheme was used to finding the solution of the Constraint estimation problem to estimate the both the correctness of the reported data and realiblity of sources. The Scheme used in that System improves the execution time of truth discovery.
As we know the Scalable and Robust Truth Discovery in Big data Social Media Sensing Application was most important thing. Daniel (Yue) Zhang, Dong Wang, Nathan Vance, Yang Zhang and Steven Mike, introduced that identifying truth information presence in noisy data was crucial task in era of big data. They recognized the problems like 'misinformation spread' and 'data sparasity' in big data social media sensing applications. They proposed the Scalable Robust Truth Discovery (SRTD) Scheme, HT Condor System and Work Framework. The SRTD scheme effectively addressed the data sparasity and misinformation challenges in big data. They evaluated the SRTD using three real world datasets. They achieved both the truth discovery accuracy and computational efficiency.
Daniel (Yue) Zhang, Yue Ma, Yang Zhang, Suwen Lin, X. Sharon Hu, Dong Wang, were addressed two important challenges conflicting interest and asymmetric and incomplete information. They develop Bottom-Up-Game-Theoretic task allocation (BGTA) framework to solve the real time and non-cooperative task allocation problem for social sensing application. They implemented a prototype of BGTA using the Nvidia Jetson boards. The result from those two real-world social sensing applications demonstrates that BGTA achieves significant performance gain in objective of applications and edge nodes.

Crowdsourcing is a process of integration acquisition and analysis of big data generated by diversity of sourced in urban spaces. Zhang Xu, Yunhuai Liu, Neil Y. Yen, describe the real time urban emergency event based on crowdsourcing using Weibo. They proposed 5W (What, Where, When, Who, Why) model, which is used to detect and describe the real time urban emergency event. The spatial and temporal information from the social media are extracted to detect real time event. They also evaluated with extensive case studies based on real urban emergency events. Model proposed by that system is applied into management field which provide useful information analyse resist urban events.

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

Daniel (Yue) Zhang, Dong Wang, Hao Zhang, Qi Li, Yang Zhang, develop a new Point-Of-Interest prediction scheme by using two core Natural Language Processing (NLP) models (Ngram and PLSA). In first, they implement a Temporal Adaptive-Ngram Model which captures the dynamic dependency between check-in points and in second Probabilistic Latent semantic Analysis (PLSA) predict to incorporate the contextual information. CAP-CP scheme can accurately predict the user's category performance in future.

Daniel (Yue) Zhang, Chao Zheng, Dong Thain, Chao Wang, Doug Thain, Chao Huang, they Introduced three fundamental challenges was dyanamic truth,Scalability and heterogeneity of streaming data.This System used effective scheme Scalable Streaming Truth Discovery(SSTD).The SSTD scheme addressed dyanamic truth challenge by explicitly modelling truth transition by using HMM based model.SSTD also effectively introduced the heterogenecity of the streaming data by integrating a feedback controller for dyanamic task allocation and resource management. To extend above work they had to explore real-time optimization (RTO) technique.

Jize Zhang, Dong Wang, presented a new analytical approach to solve the duplicate report detection problem in crowdsensing baqsed on urban issue reporting system. The fully unsupervised binary classification approach developed based on the Expectation Maximization (EM) framework. The solution evaluated by that system useful in both synthetic and real world datasets collected from smart city applications. The performance of that system improves the duplicate report detection accuracy compare to the state-of-the-art baseline. They made perfect duplicate report detection in urban crowdsensing application with the help of EM algorithm.

Xiaoxin Yin, Jiawei Han, Senior Member, they introduce and formulate the Veracity problem, which aims at resolving conflicting facts from multiple websites and finding the true facts among them.They propose TRUTHFINDER, an approach that utilizes the interdependency between website trustworthiness and fact confidence to find trustable websites and true facts. They were found TRUTHFINDER achieves highly accurate finding true facts and at the same time identifies websites which provide more accurate information.

Bo Zhao, Benjamin I. P. Rubinstein, Jim Gemmell, Jiawei Han, they experiment on two real world datasets demonstrate the clear advantage of method over the state-of-the-art truth finding methods. A case-study of source quality predicted by our model also verifies our intuition that two aspects of source quality should be considered. An efficient inference algorithm based on collapsed Gibbs sampling is developed, which is shown through experiments to converge quickly and cost linear time with regard to data size. Additionally, the method can naturally incorporate various prior knowledge about the distribution of truth or quality of sources, and it can be employed in an online streaming setting for incremental truth finding, which they prove to be much more efficient than and as effective as batch inference.

Sagar Bhuta, AvitDoshi, Uehit Doshi, Meera Narvekar, they has been found out that a number of techniques can be used to perform sentiment analysis of text. But the methods are domain specific. Moreover the techniques need to be adapted to the source from which the data is extracted. If the source is a social networking website, the language use and specific conventions need to be addressed.

Dong Wang, Lance Kaplan, Hieu Le, Tarek Abdelzaher, they described a maximum likelihood estimation approach to accurately discover the truth in social sensing applications. The approach can determine the correctness of reported observations given only the measurements sent without knowing the trustworthiness of participants. The optimal solution is obtained by solving an expectation maximization problem and can directly lead to an analytically founded quantification of the correctness of measurements as well as the reliability of participants.

Robin Wentao Ouyang, Lance M. Kaplan, Alice Toniolo, Mani Srivastava, and Timothy J. Norman, they propose new parallel and streaming truth discovery algorithms for quantitative crowdsourcing applications involving big or streaming data. Through extensive experiments, they demonstrate that both algorithms are effective. Moreover, the parallel algorithm can efficiently perform truth discovery on large datasets, and the streaming algorithm can efficiently perform truth discovery both on large datasets and in data streams. They can thus support effective and scalable truth discovery in large-scale quantitative crowdsourcing applications.

III. PROPOSED SYSTEM

In proposed system, the process of addressing misinformation spread and data sparsity in truth discovery on social media. Before classification, a classifier that contains the knowledge structure should be trained with the prelabeled tweets. After the classification model gains the knowledge structure of the training data, it can be used to predict a new incoming tweet. The whole process consists of two steps: learning and classifying. Features of tweets will be extracted and formatted as a vector. The class labels i.e. spam and non-spam could be get via some other approaches. Features and class label will be combined as one instance for training. One training tweet can then be represented by a pair containing one feature vector, which represents a tweet, and the expected result, and the training set is the vector. The training set is the input of machine learning algorithm, the classification model will be built after training process. In the classifying process, timely captured tweets will be labelled by the trained classification model.

A. Architecture

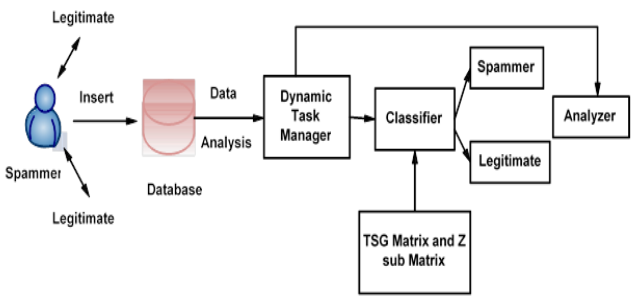


Fig 1. System Architecture

We design the architecture which uses the different legitimates and spammers which can be the input to the honeypots system. The honeypot can be send this tweets to database and then this tweets are managed by using dynamic task manager. Then we can classify the managed tweets by using naive bayes algorithm which is the most accurate classi_cation algorithm to classify the spam tweets and legitimate tweets. For the classi_er we can use the TSC matrix and sub matrix. And in last step we can analyze the spam tweets and legitimate tweets by using the analyzer.

IV. Different methodologies used in praposed system:

1. Naive Bayes algorithm:- Naive Bayes is a simple but surprisingly powerful algorithm for predictive modeling. Naive Bayes, which can be extremely fast relative to other classication algorithms. It works on Bayes theorem of probability to predict the class of unknown data set.
2. Honeypot system:- In computer terminology, a honeypot is a computer security mechanism set to detect, deect, or, in some manner, counteract attempts atunauthorized use of information systems. Generally, a honeypot consists of data (for example, in a network site) that appears to be a legitimate part of the site, but is actually isolated and monitored, and that seems to contain information or a resource of value to attackers, who are then blocked.
3. SRTD algorithm:- It is used to solve the truth discovery problem in big data social media sensing applications.It is nothing but the Scalable And Robust Truth Discovery Algorithm which is used to _nd the truth of tweets dynamically.

V. RESULT AND DISCUSSION

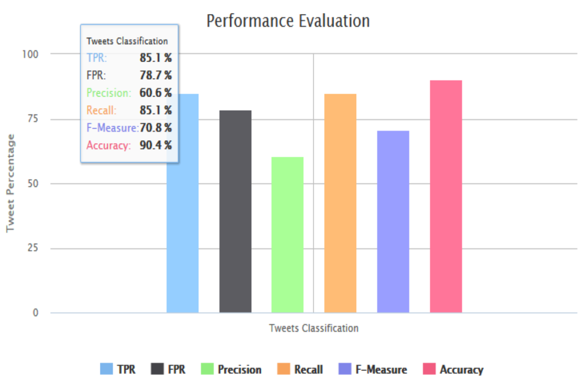


Fig 2. Analysis Graph

Parameters	Percentage
TPR	85.1
FPR	78.7
Precision	60.6
Recall	85.1
F-Measure	78.8
Accuracy	94.4

Table 1. Comparative table

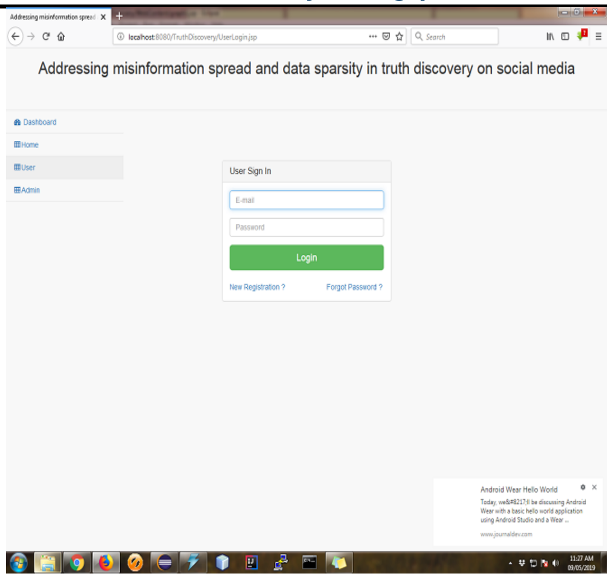
Conclusion

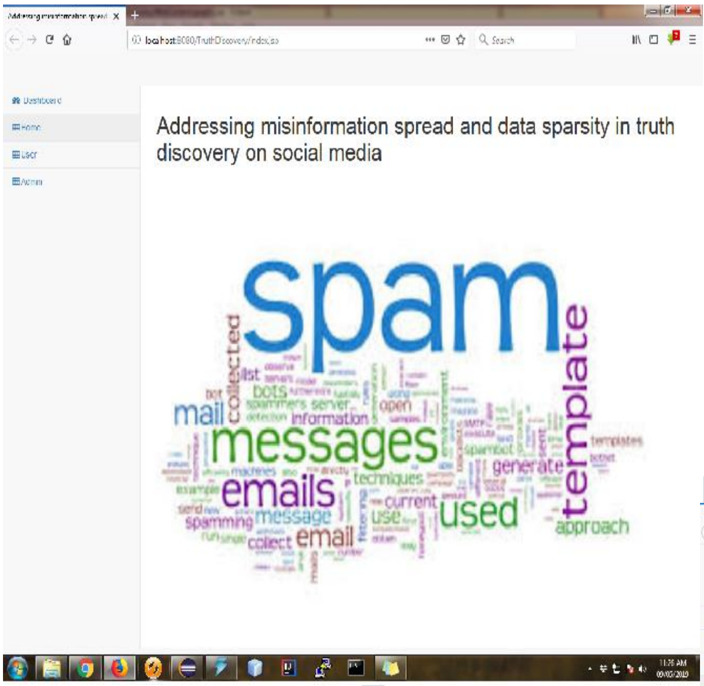
They had design and implement distributed framework using Work Queue and the HTCondor system to address the scalability challenge of the problem. In the given solutions they explicitly consider the source reliability, source credibility. They evaluated the SRTD scheme using three real world data traces collected from Twitter. The empirical results showed our solution achieved significant performance gains on both to detect dynamic truth discovery and frequently occurred information in social media sensing application. The performance of that system was gains both, to detect dynamic truth discovery and frequently occurred information in social media sensing application for noisy and sparse data.

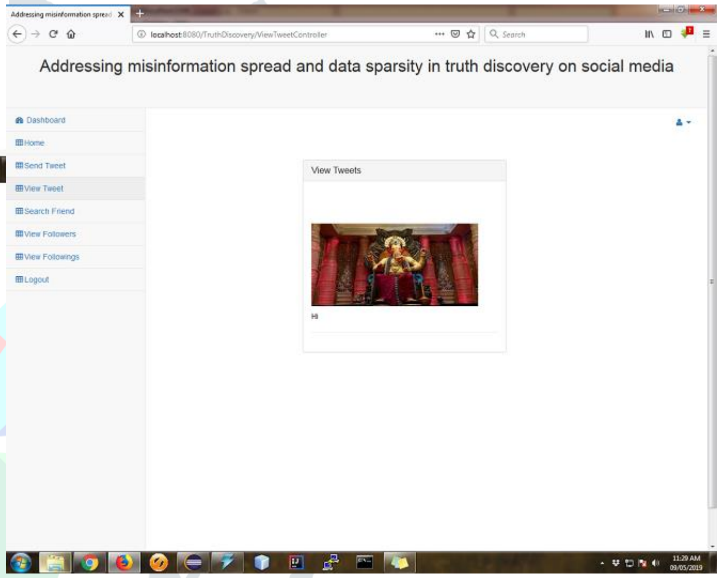
Acknowledgement

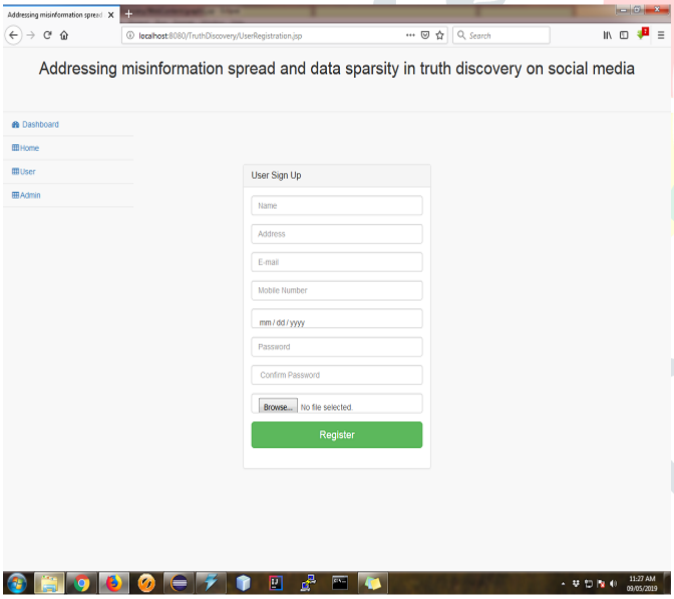
This material is based upon work supported by the National Science Foundation under Grant No. CBET-1637251, CNS-1566465 and IIS-1447795, Google Faculty Research Award 2017, and Army Research Office under Grant W911NF-17-1-0409. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

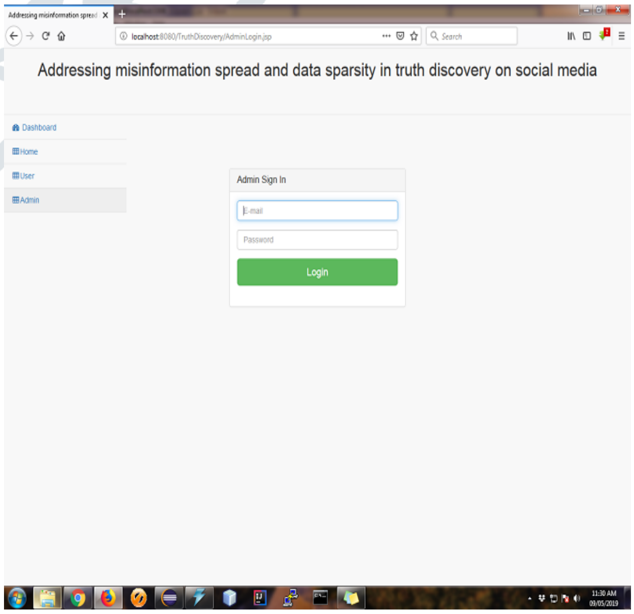
RESULT

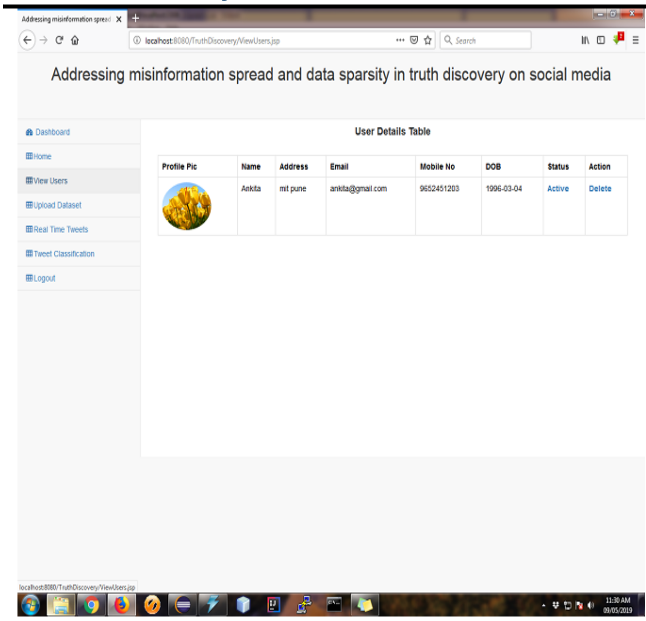

Addressing misinformation spread and data sparsity in truth discovery on social media
Dashboard
Home
User
Admin
User Sign In
Login
New Registration ?
Forgot Password ?

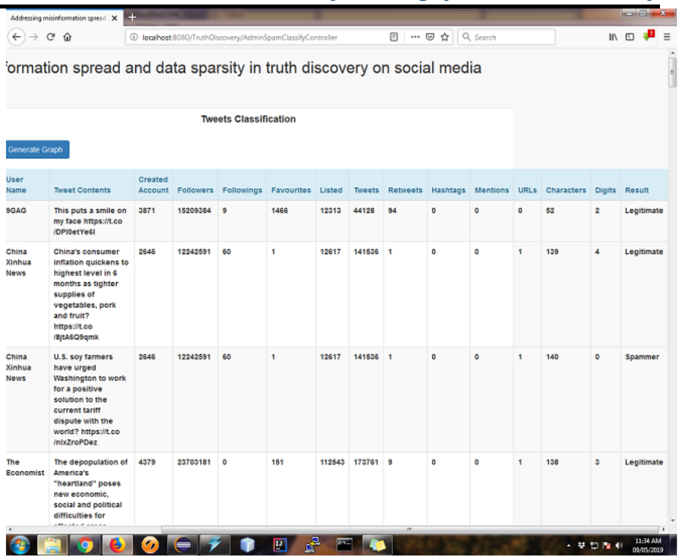

Addressing misinformation spread and data sparsity in truth discovery on social media
Dashboard
Home
User
Admin
spam
messages
emails
template
used
mail
email
collected
approach
generate

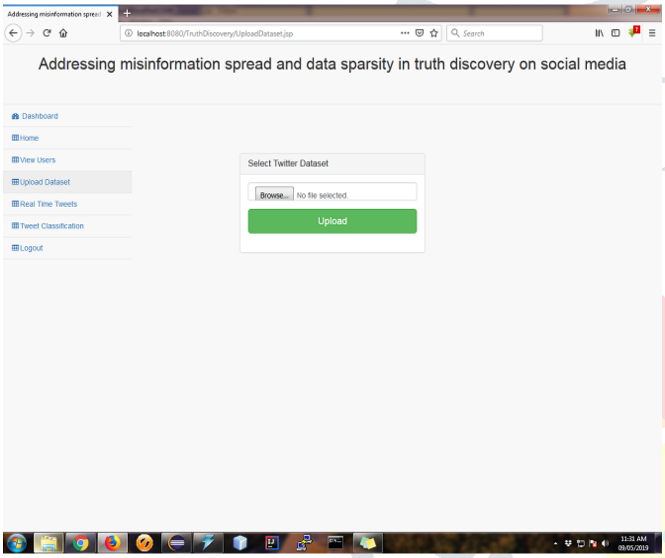

Addressing misinformation spread and data sparsity in truth discovery on social media
Dashboard
Home
Send Tweet
View Tweet
Search Friend
View Followers
View Followings
Logout
View Tweets

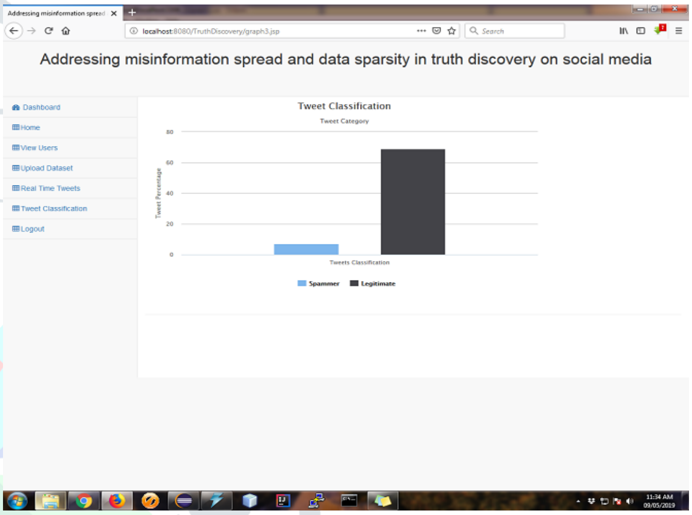

Addressing misinformation spread and data sparsity in truth discovery on social media
Dashboard
Home
User
Admin
User Sign Up
Register

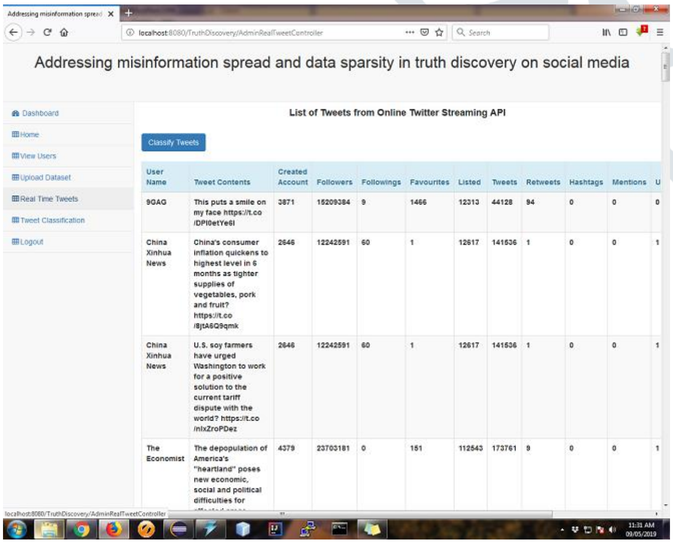

Addressing misinformation spread and data sparsity in truth discovery on social media
Dashboard
Home
User
Admin
Admin Sign In
Login

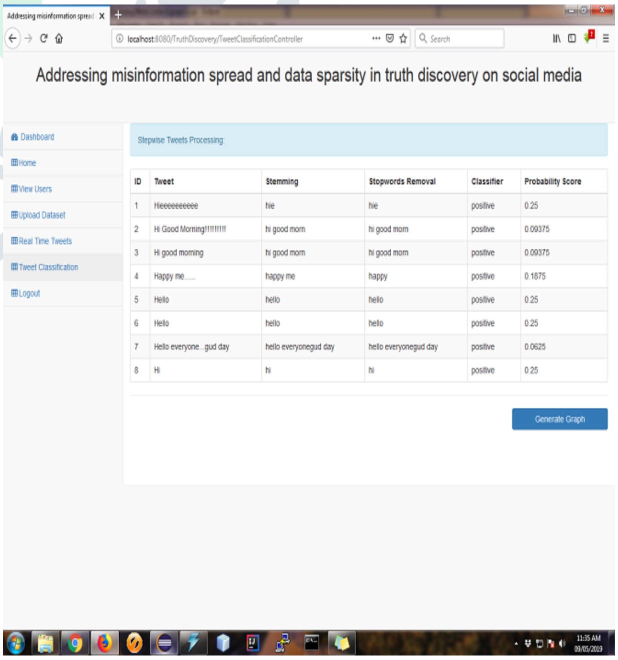

Addressing misinformation spread and data sparsity in truth discovery on social media
Dashboard
Home
View Users
Upload Dataset
Real Time Tweets
Tweet Classification
Logout
User Details Table

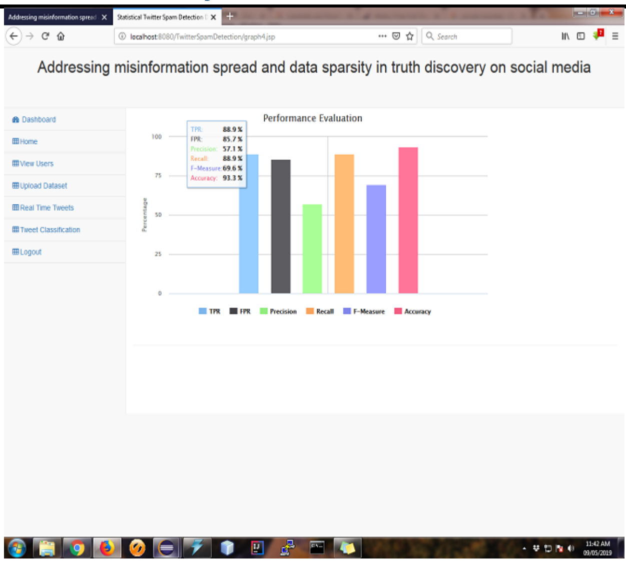

ormation spread and data sparsity in truth discovery on social media
Tweets Classification
Generate Graph


Addressing misinformation spread and data sparsity in truth discovery on social media
Dashboard
Home
View Users
Upload Dataset
Real Time Tweets
Tweet Classification
Logout
Select Twitter Dataset
Upload


Addressing misinformation spread and data sparsity in truth discovery on social media
Dashboard
Home
View Users
Upload Dataset
Real Time Tweets
Tweet Classification
Logout
Tweet Classification
Spammer
Legitimate


Addressing misinformation spread and data sparsity in truth discovery on social media
Dashboard
Home
View Users
Upload Dataset
Real Time Tweets
Tweet Classification
Logout
List of Tweets from Online Twitter Streaming API
Classify Tweets


Addressing misinformation spread and data sparsity in truth discovery on social media
Dashboard
Home
View Users
Upload Dataset
Real Time Tweets
Tweet Classification
Logout
Stepwise Tweets Processing
Generate Graph



REFERENCES

[1]Daniel (Yue) Zhang, Dong Wang, Nathan Vance, Yang Zhang and Steven Mike. ”On Scalable and Robust Truth Discovery in big Data Social Media Sensing Application.” IEEE Transction 2018

[2]Daniel (Yue) Zhang, Yue Ma, Yang Zhang, Suwen Lin, X. Sharon Hu, and Dong Wang. ”A Real-Time and Non-Cooperative Task Allocation Framework for Social Sensing Applications in Edge Computing Systems.” Department of Computer Science and Engineering University of Notre Dame IEEE RTAS 2018.

[3]Daniel (Yue) Zhang, DongWang, Yang Zhang. ”Constraint-Aware Dynamic Truth Discovery in Big Data Social Media Sensing”. Department of Computer Science and Engineering University of Notre Dame Notre Dame, IEEE 2017.

[4]Daniel (Yue) Zhang, Chao Zheng, Dong Wang, Doug Thain, Chao Huang, Xin Mu, Greg Madey .”Towards Scalable and Dynamic Social Sensing Using A DistributedComputing Framework.”Department of Computer Science and EngineeringDepartment of Aerospace and Mechanical Engineering University of Notre Dame Notre Dame, IN, USA IEEE 2017.

[5]Daniel (Yue) Zhang, Dong Wang, Hao Zheng, Xin Mu, Qi Li , Yang Zhang.” Large-scale Point-of-Interest Category Prediction Using Natural Language Processing Models.” Department of Computer Science and Engineering Department of Aerospace and Mechanical Engineering University of Notre Dame Notre Dame, IN, USA IEEE 2017.

[6]Daniel (Yue) Zhang, Rungang Han, Dong Wang, Chao Huang.” On Robust Truth Discovery in Sparse Social Media Sensing.” Department of Computer Science and Engineering University of Notre Dame Notre Dame, IN, USA IEEE 2016.

[7] Zheng Xu, Member, IEEE, Yunhuai Liu, Member, IEEE, Neil Y. Yen, Member, IEEE, Lin Mei, Member, IEEE, Xiangfeng Luo, Member, IEEE, Xiao Wei, and Chuanping Hu, Member.”Crowdsourcing based Description of Urban Emergency Events using Social Media Big Data . IEEE 2016.

[8]Jize Zhang Department of Civil and Environmental Engineering and Earth Sciences University of Notre Dame Notre Dame, US , Dong Wang Department of Computer Science and Engineering University of Notre Dame Notre Dame, US. ”Duplicate Report Detection in Urban Crowd sensing Applications for Smart City.” IEEE 2015.

[9]Jingwen Bian, Yang Yang, Hanwang Zhang, Tat-Seng Chua. ”Multimedia Summarization for Social Events in Microblog Stream.” IEEE 2015.

[10]P.T.Chen, F.Chen, Z.Qian. ”Road Trafc Congestion Monitoring in Social Media with Hinge-Loss Markov Random Fields.” IEEE 2014.

[11] Xiaoxin Yin, Jiawei Han, Senior Member and Philip S. Yu, Fellow. “Truth Discovery with Multiple Conflicting Information Providers on the Web.”IEEE 2008.

[12] Bo Zhao Benjamin I. P. Rubinstein, Jim Gemmell Jiawei Han Department of Computer Science, University of Illinois, Urbana, IL, USA.” A Bayesian Approach to Discovering Truth from Conflicting Sources for Data Integration.”

[13] Sagar Bhuta, AvitDoshi, Uehit Doshi, Meera Narvekar.” A Review of Techniques for Sentiment Analysis Of Twitter Data.”ICICT Transaction 2014.

[14] Dong Wang, Lance Kaplan, Hieu Le, Tarek Abdelzaher.” On Truth Discovery in Social Sensing: A Maximum Likelihood Estimation Approach.” Department of Computer Science, University of Illinois at Urbana Champaign, Urbana, IL 61801.

[15] Robin Wentao Ouyang, Lance M. Kaplan, Alice Toniolo, Mani Srivastava, and Timothy J. Norman. ‘’Parallel and Streaming Truth Discovery in Large-Scale Quantitative Crowdsourcing ’’IEEE 2014.