International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 03 | Mar 2019 www.irjet.net p-ISSN: 2395-0072

Big Data Driven Information Diffusion Analytics and Control on Social Network

Prof. Mahvash Khan[1], Miss. Kalyani Balapure[2], Miss. Shweta Kalambe[3], Miss. Nikita Shembekar[4], Miss. Ankita Tiple[5]

[1]*Professor, Dept. of Computer Science and Engineering, Nagpur Institute of Technology, Nagpur, Maharashtra, India*
[2,3,4,5]*Student, Dept. of Computer Science and Engineering, Nagpur Institute of Technology, Nagpur, Maharashtra, India*

Abstract - *In this paper, we are dealing with information diffusion on social media and control over it. Now a day's social media becomes most popular way to express the thoughts so that social media becomes important part of people's daily routine. They invest most of time on social media. Usually people post their news which gets replicated by their followers and friends on social media. Such post contains blogs, images, text messages and so these information gets diffused day by day however they don't know about the news truth which they are reposting. If the news is fake, then also it gets reposted by people such scenario can cause misunderstanding on people's mind they follow that news. News may be in support of something or may be in against of something or it may be related to the political parties' promotions. In such cases they misuse the people views and reaction on that news to make in favor of them. So it may lead to create big fight in opposition parties. Also this can create a difference between hearts of the people belonging to different religions. To abolish such things, we have implemented this paper. In this paper we have given solution implemented by applying some techniques and methods.*

Key Words: **Information Diffusion, sentiment analysis, naïve Bayes, Deep Learning, Linear Regression, GBTs, SVM, Diffusion Control**

1. INTRODUCTION

This paper presents the work related to the control on diffused information. The proposed approach includes extraction of data then find out the diffused then perform sentiment-analysis on the diffused information then control the over the fake information which is spreading as fast as over the social media. The proposed system consists two parts which are explained as below:

In first part, a web application, publicly are accessible through any modern browser. It will contain links to each of the dynamic real-time sentiment visualizations, as well as a search page where the user can enter a custom topic, and fresh tweets will be fetched, analyzed and rendered. In Second part, a series of open source package do specific tasks (including the sentiment-analysis module, a fetch-tweets module, Geo-lookup etc.) Each of these will be tested, documented and then publicly published for other developers to make use of. For the second part, among all data some of the diffused data will be detected and we will control on diffused data by using some techniques. So that this diffused fake data will not get propagated. Different experimentations are carried out. A database is extracted, diffused data is identified and also we will get control on diffused unwanted information by using some techniques and methodology.

1.1 Information Diffusion:

Every Information diffusion begins from specific source nodes. Let, suppose that Sara Ali Khan is a Bollywood actress promotes her movie on social media such as twitter. It may lead to take place great conversation among her fan following, and thus she is the source to start this information diffusion. Every propagator able to retrieve the data from its alongside. In previous instance Rima is a fan of Sara Ali Khan then she can only read Sara's posts and her follower's posts on her post [4]. Opine of fans on social media may be dependent on piloted situation, and social media gives a platform for people to convey their thoughts. The propagation of informatics that affects some contents can cause large consequences on our society [2]. It is accentuation to abolished the real world hypothesize since we wanted to measure information diffusion of genuine and not genuine news, which consists each diffusion beginner that accumulate data would available for us. Thus, it idealistic to remove from consideration exterior personal belonging of information diffusion through terminologies [3].

1.2 Diffusion Control:

The increasing vogue of the online social media doesn't mean that it is secure and reliable. On the conflict, the virus diverse and the confidential information diffusion have made it being an enormous headache for IT admins and people [4]. For example, "KooFace" is a Trojan Worm on Facebook, which diverse by leaving a comment on profile pages of the victim's colleague to catch a click on the malicious link . Most of system admins tensed that their workers will repost very much confidential information online. So as time passes by, it being more and more accentuate and urgent to control the virus diverse and the confidential information diffusion in

online social media. In social media there is huge amount of information are getting propagated. In this information it contains fake information as well as truthful information. In this chapter we are controlling diffused fake information by using parameters as source of the information i.e. by whom this information is spreading or who has posted this information. According to this we will decide the news is trustworthy or not. So, we will discard that information from the database. Then we will get the truthful or normal information.

2. Literature Review:

- "Weak Ties: Subtle Role in the Information Diffusion in Online Social Networks": This paper presents that weak ties play a fine role in the information diffusion in online social media. On one hand, they play a role of aqueduct, which link scheduled groups and snap through the catching of information in local areas. On the other hand, selecting weak ties advantageous to release cannot make the information diffuse speedy in the network. For possible applications, they believe that the weak ties might be of use in the domination of the virus diverse and the confidential information diffusion.

- "Information Diffusion in Online Social Networks: A Survey": In this paper they provide programs for collecting twitter data and analysis on data. They observed that real news is diverse faster than fake news on an average as well overall.

- "A Survey on Information Diffusion in Online Social Networks: Models and Methods": The paper presents the evaluation of leader of the topic and from where the news id spreading.

- "Real-time Analysis of Information Diffusion in Social Media": This paper presents that how the information gates diffused on social media and gives real-time analysis of information.

3. PROPOSED SYSTEM:

Figure shows the schematic overview of the proposed approach. The working of the proposed approach is divided into three main steps, extraction of database from social media, identifying diffused data and normal data and control on diffused data.

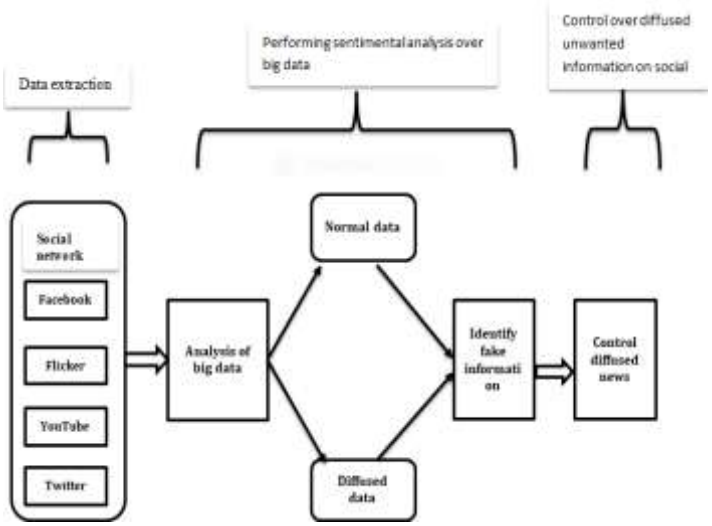


Fig: data flow diagram of data driven information diffusion analytics and control in social network

4. METHODOLOGY

There are many algorithms and methods used in practice to progress sentiment analysis systems: Hybrid, Rule-Based and Automatic. Rule-based systems incorporate a set of manually crafted rules to Perform sentiment analysis. These rules generally use a variation of inputs, classic Natural Language Processing techniques and terminologies. An example of how to develop Rule-based system is to first define a list of separated words, second is to count the number of positive words and negative words from input, final is to check if the amount of positive word appearances is greater than negative word appearances, if true, then return positive sentiment else return negative. Automatic systems use a Machine Learning thinker, exist on it text as input and return the similar classification: positive, negative or neutral (if polarity analysis is performed). The Machine Learning Classifier is implemented in typically two phases:

- The first phase involves training our model to associate input (text) to corresponding output (label or tag), which are based on test samples used for training. The connection for this process includes sending the input through a feature extraction that moves text input into a feature vector. The feature vectors go through a queue and pairs of both feature vectors and labels (positive, negative or neutral sentiment) are fed into the machine learning algorithm to create the model.

- The final phase involves prediction of sentiment score for random input. The data pipeline consists of sending raw input data through a feature extractor to be transformed into a feature vector, which is then fed to the model to generate predicted labels, such as positive, negative or neutral. Common Machine Learning Algorithms that can be used in text classification include Linear Regression,

Naive Bayes, Deep Learning, Support Vector Machines and Gradient Boosted Trees (GBTs).

- Gradient Boosted Trees (GBTs): trains a sequence of decision trees to build the classification model, uses the current group of feature vectors to predict the polarity of the label, as like negative, neutral or positive of each training instance and finally compares the prediction against the true label.

- Deep Learning: tries to imitate how the human brain functions by using artificial neural networks to process data. Sentiment analysis can be implemented to classify text in two ways, the first way is to use supervised learning if there is enough training data, else use unsupervised training followed by a supervised classifier to train a deep neural network model. Deep learning neural networks used for building sentiment classification models include recurrent neural networks, paragraph vectors, word2vec, recursive neural networks, etc.

- Support Vector Machines: are non-probabilistic models that shows text examples as points in a multidimensional space. These text examples are categorizing in different ways, such as sentiments: happy or sad, angry or pleased, etc. These categories belong to distinct regions of that multidimensional space. New texts are mapped to the same space and predicted to belong to a certain category.

- Linear Regression: is an algorithm used in statistics and machine learning to predict some value (Y) given the set of features (X).

- Naive Bayes: are set of supervised machine learning algorithms that use Bayes' theorem to predict category of text.

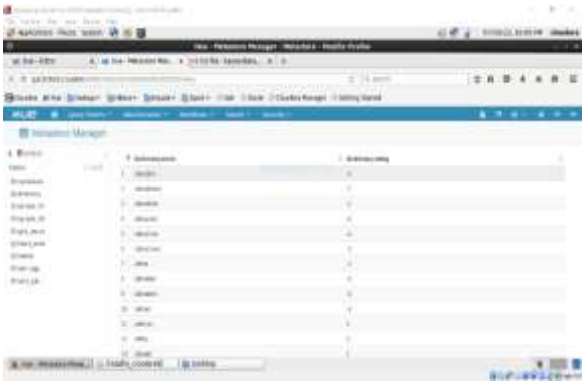
5. FUTURE SCOPE

In this accentuate future scope is that we can provide a service to various social networking sites as we are sorting only truthful information on social media. So, we can provide this system to company owner and they will make use of it to make their user free from fake news.

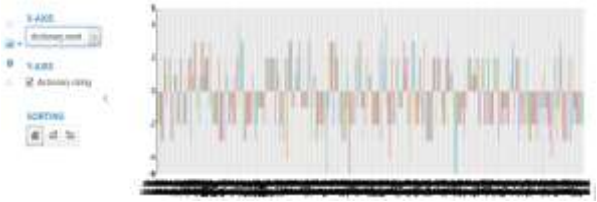
6. SCREENSHOTS

1. In this screenshot we have written querry by using querry editor in Hive.

2. fig. shows result of above query which analyses positive and negative ratings of tweets.



3. The below graph is plotted against dictionary. Word and dictionary. Rating which shows positive rating and negative rating of tweet words.



7. APPLICATION

This system is used in online social media such as Twitter, Facebook, YouTube etc. The system will provide us the data which consists only truthful data because some operations are performed on data to obtain truthful data on social media. So, users will not get fake information on social media. This system is very important to implement in various social media.

8. CONCLUSIONS

Over recent years, a variety of social networks have seen for individuals to keep in contact with others conveniently. Huge amounts of data can therefore be builds in these social media that can tell us who the very accentuate person is, who a topic leader will be, why an eventuality will spread in a specific way, and other accentuate things. To solve these matter requires multifaceted research that draws on the

fields of not only computer science, but also sociology, psychology, economics, and others.

We experimented with a set of tweets and a set of news media articles. For example, we detected influence functions of many websites and identify that they very based on the type of the website and the topic of the information. Moreover, we also perceived that the emulation and newness have a robust `force on the transformation of short document phrases in online news media. As the adoption of short, news-related document phrases become visible to be highly ruled by the influence of the few of media websites, the transformation of Twitter hashtags is ruled by a much larger set of live people, every one of which has relationally less impact. Furthermore, we also perceived that people with the massive fan followers are not the most effective in generated.

A specific exciting field for real-time survey is information diffusion, scrutinizing and predicting how information spreads. Research to understand and used information diffusion follows commonly two major directions with little interchange: On the one hand, there are knowledgeable schemas that describe and relatively information flow in social media. In this we have controlled the diffused information from social media and the users will get the truthful information so that it does not cause confusion on people's mind and their mind will not be distracted on fake things.

REFERENCES

1) Divyakant Agrawal,Basssam Bamiech,Ceren Budak,Amr EI Andrew Flanagin, and Patterson"Data-Driven Modelling and Analsis of Online Social Network"for international journal, 17 March 2011

2) Xiang Wang Kai Gao and Shashan Zhang, "A Survey on Information Diffusion in Online Social Network Model and Method" proceeding on the Hebei University of Science and Technology 050018, 25 September 2017.

3) Roman Reinche,."Comparison of Diffusion of Real and Fake News in Social Network", Technische Universitat Kaiserlutern, 02 January 2018.

4) Biao Chang, Tong Xu, Qi Liu and En-Hong Chen, "Study of Information Diffusion analysis on social network and its Applications",International journal on Automation and Computing, August 2018, volume 15, 15 August 2018.

5) Mohammad Salehan and Dan J, Kim, "The effect of sentiment on information diffusion on social media", Twenty-first Americas Conference on Information Systems, Puerto Rico, 2015.

6) Io Taxidou, "Real time Analysis of Information Diffusion in Social media", for international journal, 2014.

7) Reza Zafarani, Mohammad Ali Abbasi, Huan Liu, "Information Diffusion in Social Media", Cambridge University press, 20 April 2014.

8) Jaewon Yang, Jure Leskovec, "Modeling Information Diffusion in implicit network", Stenford University, 2017.

9) Valerio Arnaboldi, Marco Conti, Andrea Passarella, Robin LM, "Online Social network and Information Diffusion" universiry of oxford, 15 feb 2017.

10) Chunxaio Jiang, Yan Chen, K.J. Ray Lue, "Evolutionary Dynamics of Information Diffusion Over Social Network", IEEE transactions on signal processing, vol. 64, No.17, 1 September 2014.