

A RELIABILITY SYSTEM FOR FILTERING MALICIOUS INFORMATION ON SOCIAL NETWORK

Asma Nooren P, Chandana L V, Divya G A, Saniya Sadaf M
Final year B.E Students(UG)

Guided by: Mrs.Vinodha k, Assistant Professor Department of ISE, TOCE
The Oxford College of Engineering, Bengaluru ,India

Abstract : The role of social network is to enable individuals to simultaneously share information with their peers. The communities meet in person or share ideas and experiences over internet. Unfortunately the social network is misused to spread malicious information which adversely effects the society.This malicious information affects the sentiments of people and also the reputation of social network. For example during 2016 election, voting polls were falsely predicted and spread all over social media but later due to false predictions it deeply affected the sentiments of the political parties. In this work, we are focusing on fastest growing social media profoundly known as twitter. To overcome the impact of the malicious information, in our work we propose a technique which is modelled analytically by considering reputation of social network and user experience to access, analyse and validate the information .The proposed system will be validated with respect to reliable tweets obtained which will prove that the impact of malicious information will be reduced by 24% compared to existing system.

Keywords— Reliability, Reputation, Classification, User experience, Feature-Ranking, Twitter.

1. INTRODUCTION

Information reliability on Twitter has been a trending topic among researchers in the fields of both computer and social sciences, due to the recent evolution of this platform as a tool for information dissemination. Twitter enables to transfer the information in a cost-effective manner. It has now become the source of news among variety of users around the globe.

The main characteristics feature of this platform is to deliver the content in a tailored manner which allows the users to obtain news regarding their topics of choice. The development of various techniques to verify the information obtained from Twitter has been a challenging task. In this paper, we propose a reliability analysis system for assessing information on Twitter to prevent the rapid growth of fake or malicious information.

2. LITERATURE REVIEW

A new model for classifying social media users according to their behaviours.

M. Al-Qurishi et.al[1] has proposed a new model for classifying social media users according to their behaviors. Facebook and Twitter are the most popular social media that are being used as a means of social communication and sharing thoughts, knowledge and even news. Classifying huge information from these social medias using traditional data mining classification algorithms is time consuming task which needs huge processing and memory space. The authors have proposed a new approach for classifying information in social network that can give accurate result similar to support vector machine (SVM) with less processing time and consuming less memory space compare to SVM.

A Multi-stage Credibility Analysis Model for Microblogs.

M. AlRubaian et.al[2] proposed a multistage credibility analysis model for microblogs Currently, microblogs are well-known social network, which are one of the most important sources of information. In this paper, a multi-stage credibility analysis framework is proposed to prevent the proliferation of fake or malicious information on twitter. They used Naive Bayes classifier and it is enhanced by considering the relative importance of the used features to improve the classification.

A model for recalibrating credibility in different contexts and languages - A twitter case study.

A. A. AlMansour et.al[3] intended a model for recalibrating credibility in different contexts and languages. Due to the growing dependence on the WWW User- Generated Content (UGC) as a primary source for information and news, the research on web

credibility is becoming more important and more prone to threat. So this work proposes a general model to assess information credibility on UGC different platforms, including Twitter.

A Novel Prevention Mechanism for Sybil Attack in Online Social Network.

Majed AlRubaian et al.[4] suggested Sybil Defense Techniques in Online Social Networks which is a Survey regarding the problem of malicious activities in online social networks, such as Sybil attacks and use of fake identities. In this paper, they provide a comprehensive survey of literature from 2006 to 2016 on Sybil attacks in online social networks and they have reviewed existing Sybil attacks, in the context of online social networks. then they have provided a new taxonomy of Sybil attack defense.

Interactive interfaces for complex network analysis: An information credibility perspective.

J. Schaffer et al.[5] Interactive interfaces for complex network analysis: An information credibility perspective This study reveals about the impact of visualization and interaction strategies for extracting quality information from complex networks such as microblogs. Interactive node-link graph and a novel approach are the two approaches, where content is separated into interactive lists based on data properties. These two approaches are applied to overcome the problem of extracting quality information from complex networks.

3. DESIGN OF RELIABLE SYSTEM FOR FILTERING MALICIOUS CONTENT

A reliable system for filtering malicious information on twitter incorporates four components which is as depicted in the fig 3.1 and the incorporated components are: 1)Reputation of social network 2)User acceptance level 3)Feature ranking algorithm 4)Reliable classifier engine.

These components work in an algorithmic form to access, analyse and validate information and this is passed as input to decision based threshold (Dth).If $Dth > 80$ the information is rejected being malicious otherwise the information is accepted which are the reliable tweets. These results are stored in database which can be accessed for future use.

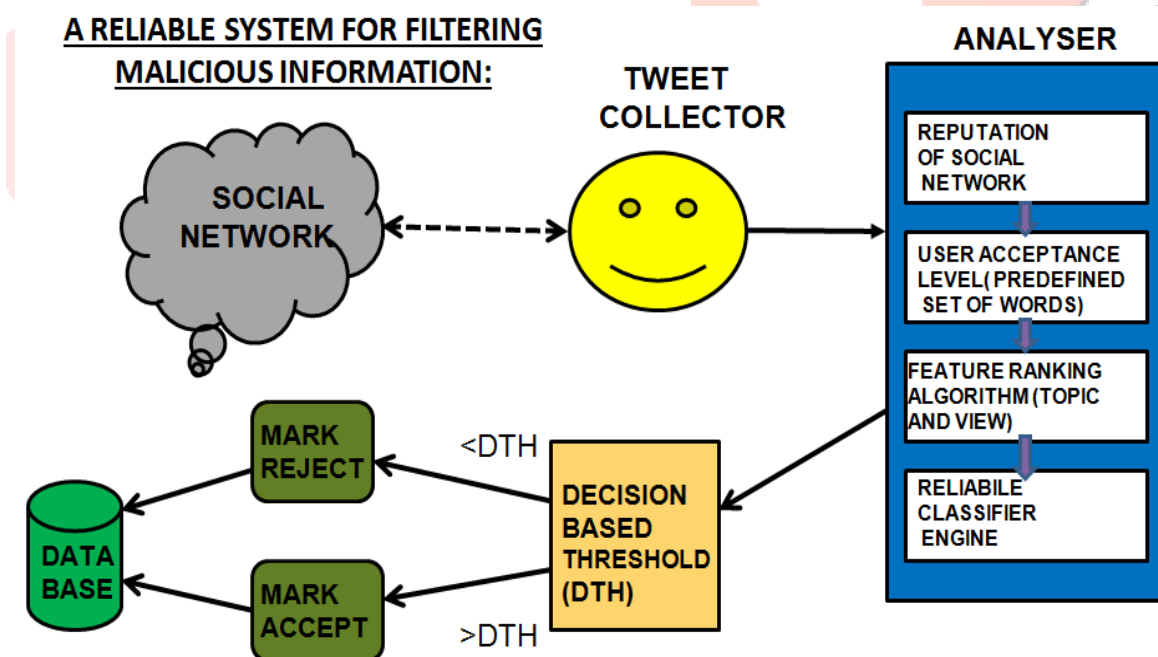


Figure 3.1: A Reliable system for filtering malicious information on twitter.

The working of reliable system involves the following units:-

Social Network:- It's a dedicated website which enables individuals to communicate with each other by posting information, comments, images, messages etc.

For example:-twitter,facebook,whatsapp,instagram

Tweet Collector:- It is a component responsible to collect the recent tweets from twitter using search and stream APIs.

Analyser :- comprises of four components which are user reputation, feature ranking, reliable classifier and user acceptance .These components work in an algorithmic form to access ,analyse and validate information collected from twitter .

Decision Based Threshold(Dth):- The results of analyser are fed to decision based component,where the threshold value is set to 80(which represents the total number of tweets collected from twitter). If $Dth > 80$ the information is rejected being malicious otherwise when $Dth < 80$ the information is accepted.

Database:- The validated results are stored in a database which can be accessed for future use.

4. ALGORITHM DESIGN AND IMPLEMENTATION

ANALYSIS OF USER RELIABILITY AND REPUTATION.

User reputation component verifies the reliability of the users and how far the information tweeted by these reputed users is trustworthy. Consequently, in this process we calculate reputation score (R) and the steps for calculating R is as followed;

Algorithm:-

Step 1: procedure CalcUserReputation (User, Tweets)

Step 2: If Tweets is empty then return 0

If User is verified then return 1

Step 3: For each $u \in \text{User}$, Calculate Users Activity Influence and Sentiment History

* UserActivity $I^p(u_i) = \sum_{u \in U, p \in T} u_i / |T|$

Where I is initial activity, p is particular topic, ui is ith user, T is tweets.

* UserInfluence $UI^{pEP} = SP^p(u_i) + I^p(u_i)$

Where SP is social popularity.

* UserSentimentHistory $\Delta_{ui} = \sum T_{ui}^+ / \sum T_{ui}^+ + \sum |T_{ui}^-|$

Where T_{ui}^+ is positive tweet of i^{th} user, T_{ui}^- is negative tweet of i^{th} user

Step 4: The reputation R is given by,

$R = \text{sentiment history (SH)} * \text{user influence (UI)}$

Step 5: End process

PRIORITIZING FEATURES BASED ON RE-TWEET COUNT.

Feature ranking component returns the reliability of tweets, i.e, it tells us how far a tweet can be trusted based on re-tweet counts. Higher the re-tweet count for a tweet more the particular tweet can be trusted. Judgement matrix for feature ranking (FR) consisting features (F) are given as follows

$$F(R) = \begin{pmatrix} 1 & f_{12} & \dots & f_{1n} \\ f_{21} & 1 & \dots & f_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ f_{n1} & f_{n2} & \dots & 1 \end{pmatrix}$$

Algorithm

Step 1:Procedure FEATURERANK(FR)

Step 2:For each column $C \in FR$

Step 3:Normalizing features is given by (S) where , $S \leftarrow \sum_{i \in c} (F_i)$ with respect to the row.

Step 4:End for.

Step 5:for each feature $F_i \in FR$

Step 6: $FR^{\wedge} \leftarrow$ normalizing FR by dividing each entry on S

Step 7:calculate a list of all the ranked features with respect to

$$RC = \left[\frac{\prod_{j=1}^n f_{ij}^{1/n}}{\left(\prod_{i=1}^n \left[\prod_{j=1}^n f_{ij} \right] \right)^{1/n}} \right]$$

Step 8:End for

Step 9:RF \leftarrow create a list of all the ranked features with respect to RC

Step 10:return RF

Step 11:end procedure.

CLASSIFICATION OF TWEETS.

The main aim of reliable classifier engine is to classify positive, negative and neutral tweets and eliminates the negative tweets. The classification is based on naïve baye's concept.

$$P(A/C) = \frac{P(C/A) P(A)}{P(C)}$$

Where,

A is feature or attribute of training data,

C is conditions applied on training data,

P(A/C) is probability of attribute based on condition.

P(C/A) is probability of condition based on attribute.

P(A) is probability of attribute.

P(C) is probability of condition.

The following function shows the classification of positive and negative tweets.

```
<p class="postweet">//for positive tweets
<%
out.print("@ " + tweet.getUser().getScreenName() + " - " + tweet.getText());
set--;
}
} else if (score <= -1) {
neg++;
if (set > 0) {
%>
</p>
<p class="negtweet">//for negative tweets
<%
out.print("@ " + tweet.getUser().getScreenName() + " - " + tweet.getText());
set--;
}
} else if ((score < 1) && (score > -1))
{
neu++;
if (set > 0) {
%>
</p>
```

USER SEARCH HISTORY ANALYSIS.

In user search history analysis we determine the topic of interest of a particular user based on their respective search history. The frequently searched topic in the user's search history will be the most reliable topic for the user by this we can obtain the set of tweets which are more trustworthy from the user point of view.

The following function is used to access user's sentiment history

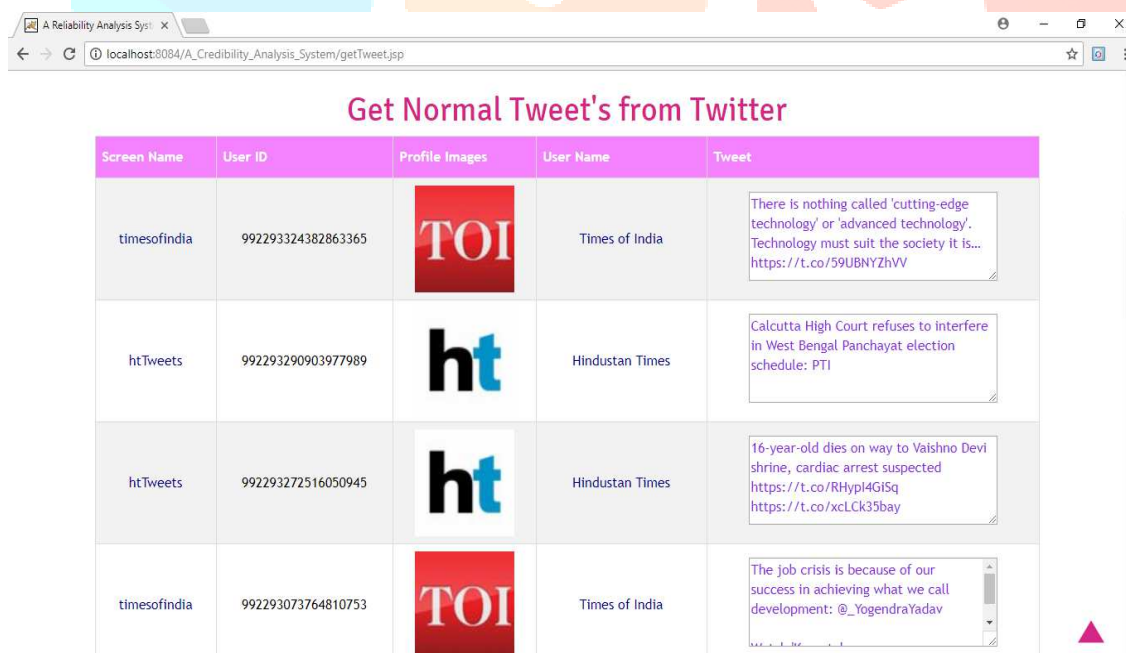
```
<center><h4>Get User Sentiment Search History </h4></center>
<div class="center btmspace-80">
<%
ConfigurationBuilder cf = new ConfigurationBuilder();
cf.setDebugEnabled(true)
.setOAuthConsumerKey("aV8lanFby7bTEMI2JXfJPiuB7")
.setOAuthConsumerSecret("3Di9ULBEzWt1PJUtCgvUnU7vXvvVE74cdxrNA7pfVeF1sTSSsty")
.setOAuthAccessToken("759307560369303553-X1kMf7u6BapUEMqQIQRMaR9fCuXgoyd")
.setOAuthAccessTokenSecret("awCfmbazBXRyk1ddMF7sUaCSD1Xkr4cYc6T7QsAncpC2g");
TwitterFactory tf = new TwitterFactory(cf.build());
twitter4j.Twitter twitter = tf.getInstance();
java.util.List<Status> status = twitter.getHomeTimeline(); %
```

5. RESULT'S DISCUSSION

We validate our system on different datasets of Twitter content, Our results show that the system which involved a reputation-based component, reliable classifier engine, user acceptance component and feature ranking algorithm provides a significant and accurate reliability assessment. The major outcomes of our proposed system is that on validating information with respect to error rate the impact of malicious information will be reduced by 24% when compared to existing system.

We are assessing the recent tweets from twitter using stream and search API.

The figure 5.1 illustrates the example of assessing the recent tweets from the twitter.







Screen Name	User ID	Profile Images	User Name	Tweet
timesofindia	992293324382863365		Times of India	There is nothing called 'cutting-edge technology' or 'advanced technology'. Technology must suit the society it is... https://t.co/59UBNYZhVV
htTweets	992293290903977989		Hindustan Times	Calcutta High Court refuses to interfere in West Bengal Panchayat election schedule: PTI
htTweets	992293272516050945		Hindustan Times	16-year-old dies on way to Vaishno Devi shrine, cardiac arrest suspected https://t.co/RHypl4GiSg https://t.co/xcLck35bay
timesofindia	992293073764810753		Times of India	The job crisis is because of our success in achieving what we call development: @_YogendraYadav

Fig 5.1 : Example for extracting recent tweets from twitter.

Here we rank the tweets based on retweet count. Higher the retweet count more the trustworthy of the tweet.

In the fig 5.2, the topic NDTV has the highest retweet count of 6. Therefore it is more reliable when compared to other tweets.

User ID	ScreenName	Profile Images	User Name	Tweet	Re Tweet Count
992292385584435202	ndtv		NDTV	#Blog: Ravish Kumar on his photo with earphones going viral	6
992292451854442496	timesofindia		Times of India	. Saugata Bhattacharya and are discussing Digital disruption: Is India ready for im...	4
992292273403527168	ndtv		NDTV	"Sidaramaiah Supported Jihadi Elements": Yogi Adityanath In Karnataka #NDTVNewsBeps	4
992292026958794753	htTweets		Hindustan Times	RT #Deadpol a Titanic, gets #CelineDion to sing theme song. Watch	3

Fig 5.2 : Features ranked based on re-tweet count.

Fig 5.3 gives the sample output of user reputation where the reliability of user is verified based on reputation score. So here the reliable user is *ht* with the highest reputation score of 795858.

User ID	ScreenName	Profile Images	User Name	Tweet	User Reputation Score
992293290903977989	htTweets		Hindustan Times	Calcuta High Court refuses to interfere in West Bengal Panchayat election schedule: PTI	7.95858
992293272516050945	htTweets		Hindustan Times	-year-old dies on way to Vaishno Devi shrine, cardiac arrest suspected	7.95858
992292924237860870	htTweets		Hindustan Times	President Xi Jinping says Marxism still 'totally correct' for China	7.95858
992292762362900480	htTweets		Hindustan Times	PC Jewelers wipes out \$2. market cap after founder gifted shares to family	7.95858

Fig 5.3: Tweets based on user reputation score.

This component classifies the positive, negative and neutral tweets and eliminates the negative tweets.

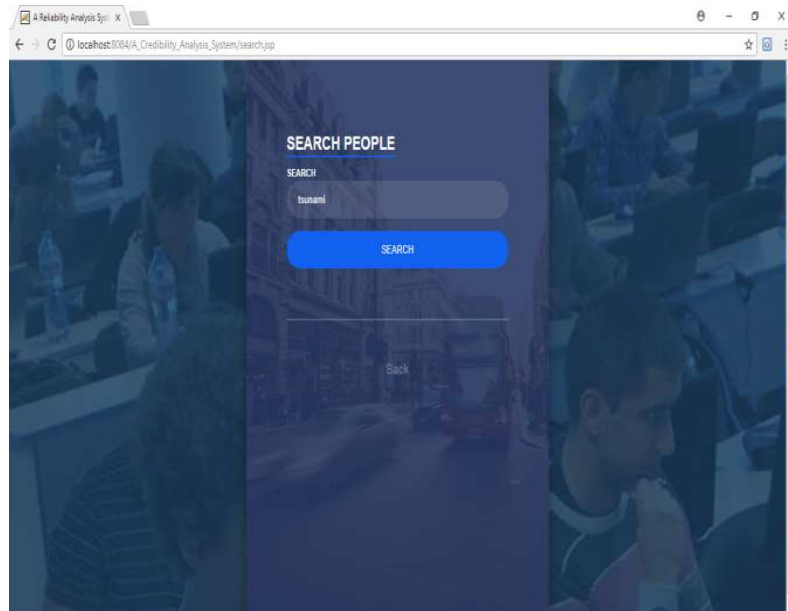


Fig 5.4: This snapshot illustrates searching of tweets

The result of the topic tsunami has 33.33% of positive and 33.33% of negative tweets as shown in the fig 5.5.

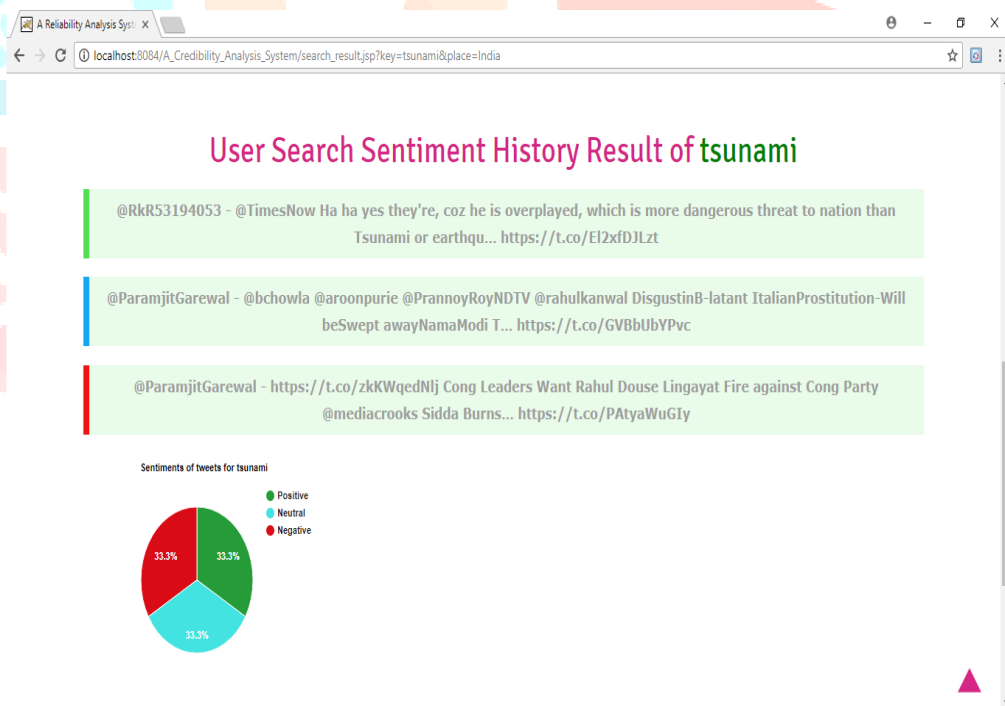


Fig 5.5: Classification of positive and negative tweets for the topic tsunami.

In fig 5.6 we notice that the negative tweets are eliminated for the tsunami

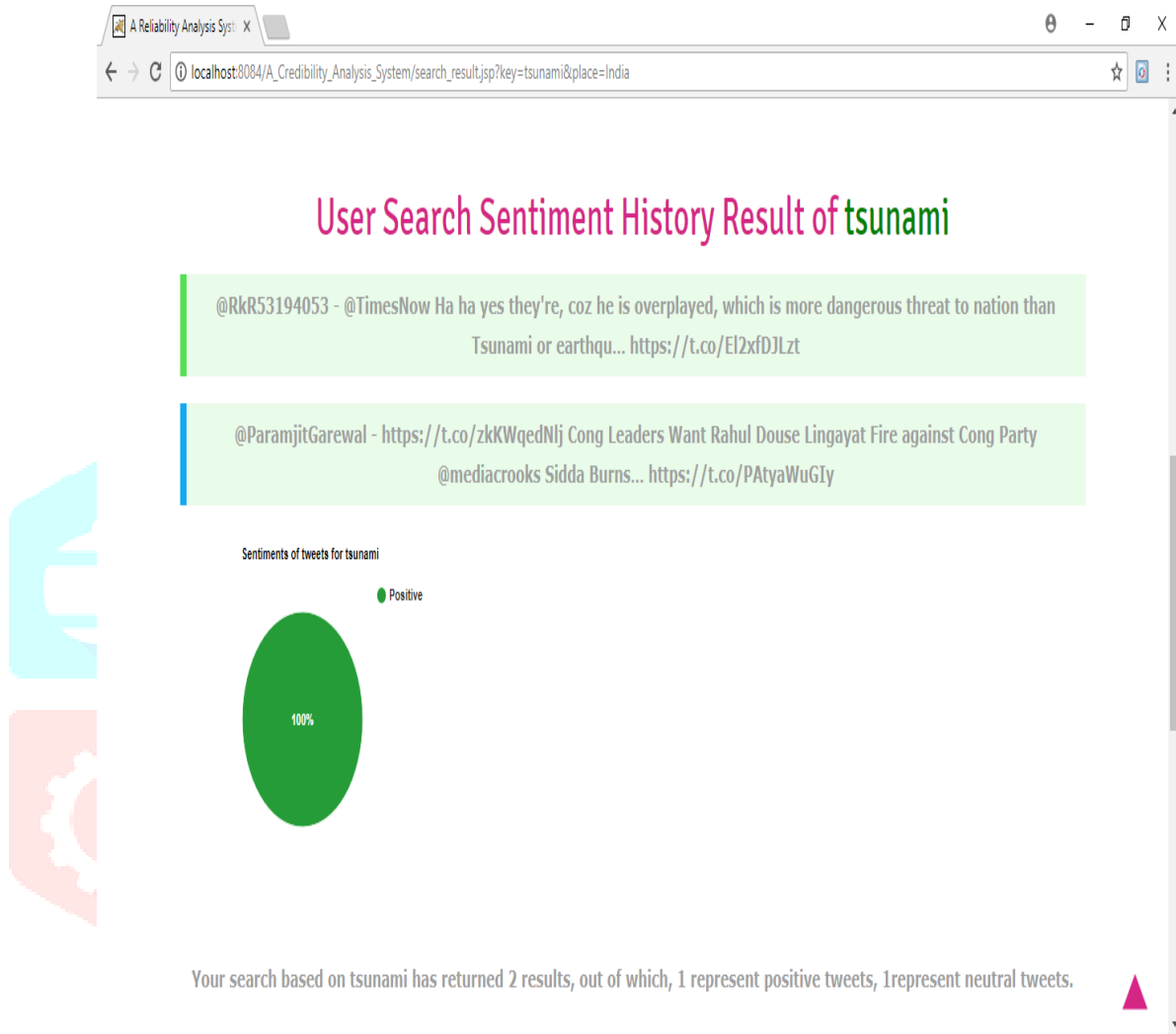


Fig 5.6: Elimination of negative tweets for the topic tsunami

Fig 5.7 gives idea about the user’s preferences based on their search history. In this example the users most search topic is csk.

User ID	User Name	State	Country	Search Keyword
7	saniya	karnataka	india	rcb
7	saniya	karnataka	india	csk
7	saniya	karnataka	india	rcb
7	saniya	karnataka	india	ramya
7	saniya	karnataka	india	sonam
7	saniya	karnataka	india	rcb
7	saniya	karnataka	india	bjp
7	saniya	karnataka	india	csk
7	saniya	karnataka	india	csk
7	saniya	karnataka	india	bjp
7	saniya	karnataka	india	csk
7	saniya	karnataka	india	csk
7	saniya	karnataka	india	csk
7	saniya	karnataka	india	csk

Fig 5.7: User’s search history.

Fig 5.8 gives the performance of reliable system for filtering malicious content which is in the form of a pie chart. This pie chart gives the comparison of proposed system with the existing system. Here we notice that the proposed system shows the greater efficiency compared to existing system.

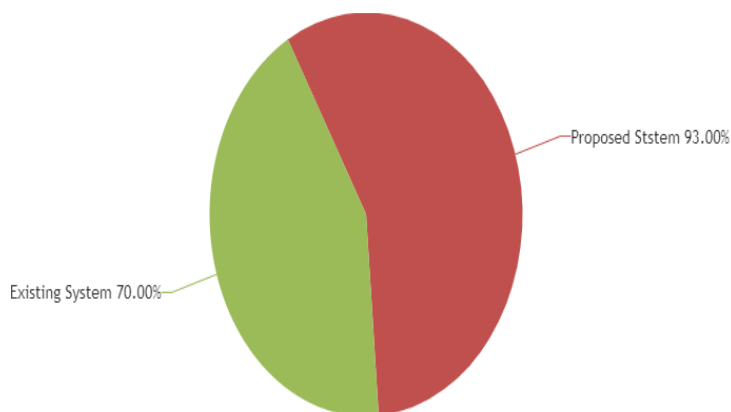


Fig 5.8:Performance of reliable system for filtering malicious content

6. CONCLUSIONS AND FUTURE WORK

Conclusions:

Social networks like twitter, facebook etc play a vital role in exchanging information among people. The information exchanged through these social networks maybe malicious and it may have an adverse affect on the sentiments of people.

To overcome the dissemination of malicious information which affects the sentiment of the people we proposed a reliable system which includes four components integrate together and work in an algorithmic form as a result of which we can observe that the error rate is reduced by 24%.

Future work:

The system can be improved and extended with the following aspects in future:

- The system which is getting implemented with four major components to reduce malicious information is best suited and possibly the most essential design which could have happened now.
- In future by the technology being changed or there are things invented everyday, there are expectations that the malicious information could be further reduced by more than 24%.
- Also in future our system can be programmed as a weapon to identify the inconsistent users who are frequently tweeting malicious information.
- The system can be implemented with different features, one of them is elimination of negative tweets prior it is being posted.

7. REFERENCES

- [1] M. Al-Qurishi, R. Aldrees, M. AlRubaian, M. Al-Rakhami, S. M. M. Rahman, and A. Alamri, "A new model for classifying social media users according to their behaviors," in Web Applications and Networking (WSWAN), 2015 2nd World Symposium on, 2015, pp.15
- [2] M. AlRubaian, M. Al-Qurishi, M. Al-Rakhami, S. M. M. Rahman, and A. Alamri, "A Multi-stage Credibility Analysis Model for Microblogs," presented at the Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015, Paris, France, 2015 .
- [3] A. A. ALMANSOUR, L. BRANKOVIC, AND C. S. ILIOPOULOS, A MODEL FOR RECALIBRATING CREDIBILITY IN DIFFERENT CONTEXTS AND LANGUAGES - A TWITTER CASE STUDY.
- [4] Majed AlRubaian, Muhammad Al-Qurishi, Sk Md Mizanur Rahman, and A. Alamri, "A Novel Prevention Mechanism for Sybil Attack in Online Social Network," presented at the WSWAN'2015, 2015.
- [5] J. Schaffer, B. Kang, T. Hollerer, H. Liu, C. Pan, S. Giyu, and J. O'Donovan, "Interactive interfaces for complex network analysis: An information credibility perspective," in Pervasive Computing and Communications Workshops (PERCOM Workshops), 2013 IEEE International Conference on, 2013, pp. 464-469.
- [6] A. A. AlMansour, L. Brankovic, and C. S. Iliopoulos, "Evaluation of credibility assessment for microblogging: models and future directions," in Proceedings of the 14th International Conference on Knowledge Technologies and Data-driven Business, 2014, p. 32.
- [7] Majed AlRubaian, Muhammad Al-Qurishi, Sk Md Mizanur Rahman, and A. Alamri, "A Novel Prevention Mechanism for Sybil Attack in Online Social Network," presented at the WSWAN'2015, 2015
- [8] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter," presented at the Proceedings of the 20th international conference on World wide web, Hyderabad, India, 2011.
- [9] S. Y. Rieh, M. R. Morris, M. J. Metzger, H. Francke, and G. Y. Jeon, "Credibility Perceptions of Content Contributors and Consumers in Social Media," 2014.
- [10] M. R. Morris, S. Counts, A. Roseway, A. Hoff, and J. Schwarz, "Tweeting is believing?: understanding microblog credibility perceptions," in Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work, 2012, pp. 441-450.
- [11] Pal, A. and Counts, S. What's in a @name? How Name Value Biases Judgment of Microblog Authors. in Proc. ICWSM, AAAI (2011)

- [12] A. Gupta, P. Kumaraguru, C. Castillo, and P. Meier, "Tweetcred: Real-time credibility assessment of content on twitter," in Social Informatics, ed: Springer, 2014, pp. 228-243
- [13] Metaxas, Panagiotis Takas, Samantha Finn, and Eni Mustafaraj. "Using TwitterTrails. com to Investigate Rumor Propagation." Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work & Social Computing. ACM, 2015
- [14] A. Gupta and P. Kumaraguru, "Credibility ranking of tweets during high impact events," in Proceedings of the 1st Workshop on Privacy and Security in Online Social Media, 2012, p. 2
- [15] M. Mendoza, B. Poblete, and C. Castillo, "Twitter Under Crisis: Can we trust what we RT?," in Proceedings of the first workshop on social media analytics, 2010, pp. 71-79
- [16] Westerman, D., Spence, P.R., and Van Der Heide, B.: 'A social network as information: The effect of system generated reports of connectedness on credibility on Twitter', Computers in Human Behavior, 2012, 28, (1), pp. 199-206
- [17] B. Kang, J. O'Donovan, and T. Höllerer, "Modeling topic specific credibility on twitter," in Proceedings of the 2012 ACM international conference on Intelligent User Interfaces, 2012, pp. 179-188
- [18] Y. Ikegami, K. Kawai, Y. Namihira, and S. Tsuruta, "Topic and Opinion Classification Based Information Credibility Analysis on Twitter," in Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on, 2013, pp. 4676-4681

