Towards Automated Real-Time Detection of Misinformation on Twitter

Suchita Jain , Vanya Sharma and Rishabh Kaushal
Department of Information Technology
Indira Gandhi Delhi Technical University for Women, Delhi, India

***Abstract*—Online Social Media (OSM) in general and more specifically micro-blogging site Twitter has outpaced the conventional news dissemination systems. It is often observed that news stories are first broken in Twitter space and then the electronic and print media take them up. However, the distributed structure and lack of moderation in Twitter compounded with the temptation of posting a news worthy story early on Twitter, makes the veracity of information (tweet) a major issue. Our work is an attempt to solve this problem by providing a approach to detect *misinformation/rumors* on Twitter in real-time automatically. We define a rumor as any information which is circulating in Twitter space and is not in agreement with the information from a credible source. For establishing credibility, our approach is based on the premise that verified News Channel accounts on Twitter would furnish more credible information as compared to the naive unverified account of user (public at large). Our approach has four key steps. Firstly, we extract live streaming tweets corresponding to Twitter trends, identify topics being talked about in each trend based on *clustering* using hashtags and then collect tweets for each topic. Secondly, we segregate the tweets for each topic based on whether its tweeter is a verified news channel or a general user. Thirdly, we calculate and compare the contextual and sentiment mismatch between tweets comprising of the same topic from verified Twitter accounts of News Channels and other unverified (general) users using *semantic* and *sentiment analysis* of the tweets. Lastly, we label the topic as a rumor based on the value of mismatch ratio, which reflects the degree of discrepancy between the news and public on that topic. Results show that a large amount of topics can be flagged as suspicious using this approach without involvement of any manual inspection. In order to validate our proposed algorithm, we implement a prototype called *The Twitter Grapevine* which targets rumor detection in the Indian domain. The prototype shows how a user can leverage this implementation to monitor the detected rumors using activity timeline, maps and tweet feed. User can also report the rumor as incorrect which can then be updated after manual inspection.**

Keywords*—Trending Rumor, Misinformation Detection, Verified News Channels

I. INTRODUCTION

With the empowerment and increasing popularity of online social networks, Twitter has evolved into a source of news for many users around the world. It has become a fast, efficient and easily accessible source for news- enthusiasts all over the globe. People are turning to Twitter to seek information about emergency situations and daily events.

Considering legacy systems, the delivery of news using old media methods like television channels, radio and print communication has been challenged by social platforms, which present a newer way to get news to people. It is faster, easily accessible and has a much wider audience on a single platform. This makes Twitter a vast improvement over the existing news dissemination systems. Due to crowd-sourced information, Twitter suffers from the problem of credibility and misinformation. In order to post information on Twitter as soon as it happens, the credibility of the content is at stake. It becomes difficult to post correct and verified information when a sense of competition arises to post tweets as and when they occur.

The spread of misinformation being conveyed to large masses is becoming a huge problem these days. Since a large number of users are exposed to the news instantaneously and they tend to believe whatever information it has to be true, an environment of anxiety is developed. For example, The Boston Marathon Bombings left millions of users believe that the US President Barack Obama was severely injured. The government and authorities had to spend a lot of money and time to recover from the instability thus caused.

The target problem is highly relevant in todays world as people especially youngsters are turning to Twitter for news. People, often are found to get swayed by the information and start believing whatever they see on Twitter. People have a right to know whether the information they are seeing is trustworthy or not. Relevance of this domain of work can also be seen by observing the severe effect of wrong information propagated on Twitter. It may be tarnishing the image of individuals and organizations, or to induce panic and instability in general public.

It is of prime importance to search for faster ways to tackle this problem in order to cut down the wastage of revenue and resources spent by the government. One such example is the rumor that Burj Khalifa was painted/ lighted in Indian Tricolor when Narendra Modi visited UAE.

It is therefore very essential to differentiate between misinformation and credible news in real time. We deal with the problem below, to approach new, practical ways to deal with misinformation on Twitter.

The rest of the paper is structured as follows. Next section indicates the related work in brief. In Section III, we describe the problem. The proposed approach is explained in section IV. Section V gives the implementation and results of the proposed system. The approach is analysed for accuracy in section VI. Finally, in Section VII and VIII, the paper is concluded and open issues are discussed.

978-1-5090-2029-4/16/$31.00 @2016 IEEE

II. LITERATURE REVIEW

A number of researches are being carried out to address the problem of credibility of information on Twitter and similar platforms in an efficient and timely manner. Research is being carried out to find new and fast methods to detect rumors and combat them in an effective manner. In this section, we discuss some of the most prominent works in the field.

In *Twitter Under Crisis: Can we trust what we RT?* [1], Marcelo Mendosa et al. have detected rumors during the Chile earthquake based on the assumption that rumors tend to be more questioned than other tweets. Kate Starbird et al. have discussed the impact of rumors during the Boston marathon bombing in their work [2]. They have examined closely the relationship between misinformation and corrections on Twitter. In *Rooting out the rumor culprit from suspects*[3], the flow of rumor has been traced using a retweet graph to find out its source. *Automatic detection and verification of rumors on Twitter* by Vosoughi, Soroush, [4], tried to solve the issue by classifying and clustering assertions made about that event through a speech-act classifier and then evaluating the veracity by examining three aspects of information spread: linguistic style used to express rumors, characteristics of people involved in propagating information, and network propagation dynamics. In *Enquiring Minds: Early Detection of Rumors in Social Media from Enquiry Posts* [5], rumors are detected using signal tweets, that is, tweets that contain skeptical enquiries: verification questions and corrections. Xing Zhou et al. have worked on real-time news certification on Chinese social network, Sina Weibo[6] by building an ensemble model that combine user-based, propagation-based and content-based model. Raveena Dayani et al. have worked to find rumors using natural language processing by maintaining a look-up table for keywords in rumors[7].

III. PROPOSED APPROACH

A. Data Collection

To detect trending rumors, we have collected tweets along with their meta data according to the following algorithm(refer to Algorithm 1).

Algorithm 1 collectTopicwiseTweets($TrendSet$)

1: **for each** $Trend \in TrendSet$ **do**
2: $TweetSet \leftarrow getTweets(Trend)$
3: **if** $containsAmbiguousTweets(TweetSet)$ **then**
4: $TopicSet \leftarrow detectTopic(TweetSet)$
5: **for each** $Topic \in TopicSet$ **do**
6: $TopicwiseTweetSet \leftarrow getTweets(Topic)$ // TopicwiseTweetSet has structure T (Topic,tweets)
7: **end for**
8: **end if**
9: **end for**
10: **return** $TopicwiseTweetSet$

The procedure for generating graphs to detect topics used in algorithm 1 is explained in the algorithm 2:

Algorithm 2 detectTopic($TweetSet$)

1: $G \leftarrow generateHashtagCooccurrenceGraph(TweetSet)$ // G has hashtags as the vertices and frequency of the co occurrence of the hashtags in a tweet as the weight of edges
2: $TopicSet \leftarrow CommunityDetection(G)$ // Each community in the graph is assumed to talk about a single topic
3: **return** $TopicSet$

Top Twitter trends are collected along with their tweets. The trends are then filtered to limit them to English language. Screening is done to limit the analysis to trends which have tweets with some amount of ambiguity. This is achieved by searching for question words in trend. Screening of trends is done based on the assumption that rumors tend to be more questioned than other information. Then trends with ambiguity are processed to find what topics they are talking about. A topic is defined as a set of hashtags characterizing what is being talked about in that particular trend. For example, ModiInUAE trend has topic characterised by hashtags #Tricolor, #IndianFlag, #BurjKhalifa. Topic detection is done in the following manner: first, we parse the tweets of the trend. Then we generate a graph of hashtags occurring in these tweets to signify their inter-relation. The vertices of this graph represent the hashtags and edge weights represent the frequency of the co-occurrence of the hashtags. Each time we encounter a pair of hashtags in a tweet, their edge weight is increased. Thus a larger edge weight would signify that the pair of hashtags has occurred together more frequently in the set of tweets. The hashtag sets that occur frequently in the set are found out by finally processing the graph by a well established community detection algorithm, walk trap in our case. Every cluster now represents a topic. This topic now contains a set of closely related hashtags that can be used for mining misinformation. A sample of such graph is shown in fig 1.

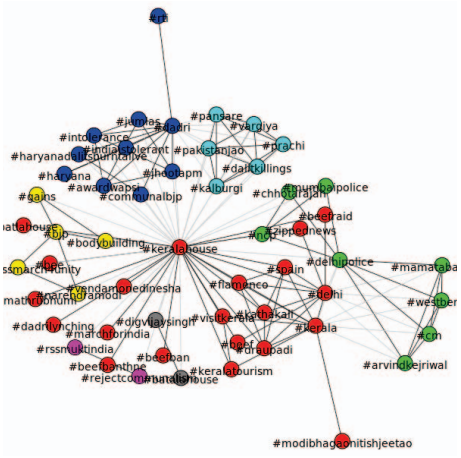

Fig. 1: A sample graph to illustrate topic detection technique. The graph is formed in the back-end for the trend *#KeralaHouse*

The graph in figure 1 has been created for the trend #KeralaHouse using the python igraph library. As a result, it gives the set of topics that the trend has generated on Twitter. Set of vertices with same color represents a cluster. The more the edge weight is, the darker it appears on the graph. Each cluster is then sent as a query to fetch tweets with its meta data from the Twitter space. The resulting tweets are passed as input to the rumor detection algorithm.

B. Rumor Detection

In our context, rumor is defined as any information posted on Twitter, that many people believe to be true, but it contrasts with the news tweets from the verified[1] news channels.

Our algorithm is based on the following premise: *"Verified News Channel accounts on Twitter would furnish credible information as compared to the naive unverified account of user."* A justification of this premise can be seen from the process used by a verified news channel to post information. News channels verify the information before posting it. They hold accountability for the information they post. They strive to maintain their reputation to post the correct news as fast as possible and consider the fact that the news would affect a large user base. The news channel accounts are further verified by Twitter for their identity, preventing fraud accounts. Thus, information from a verified news channel account can be considered trustworthy.

The flow diagram for the algorithm is given in fig 2.

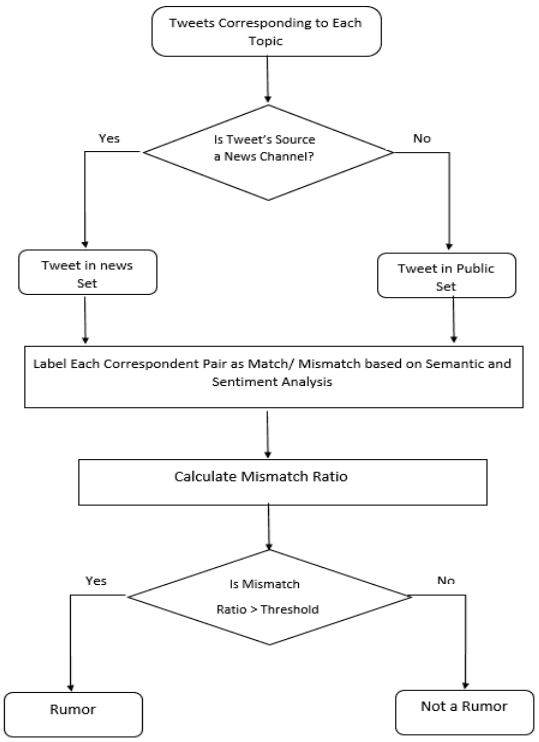

Fig. 2: Rumor detection flowchart

To detect misinformation, the tweets are segregated into two sets, namely news set and public set, based on their source being news channel accounts or otherwise. The tweets from a verified news channel are categorized as news tweets, all the other tweets fall into public tweet set.

All the tweets from the public set are then compared with the those in the news set using semantic and sentiment analysis.

[1]Verified accounts are the accounts whose credentials have been verified by Twitter. Verification is currently used to establish authenticity of identities of key individuals and brands on Twitter. Any account with a blue verified badge on their Twitter profile is a verified account.

Finally, each pair of the cross product of the public and news tweet sets is labelled as *match* or *mismatch* according to the rule engine in fig 3.

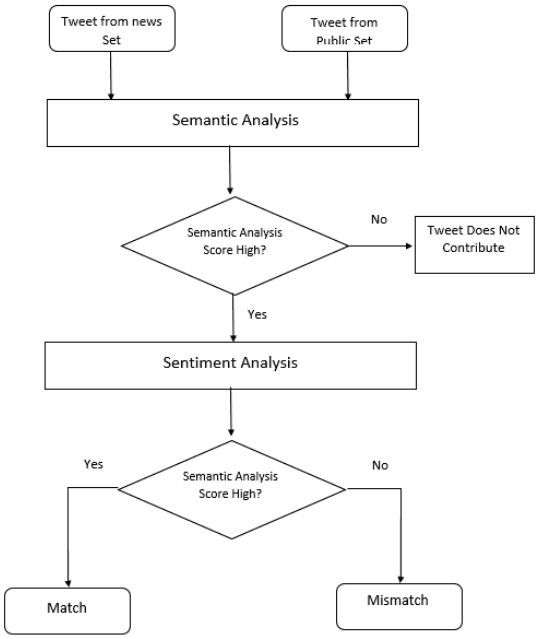

Fig. 3: Rule engine to label tweets as match or mismatch

Semantic analysis is used to determine contextual similarity between the tweets. The analysis is done in two steps. First, the stop words are removed so that a more accurate analysis of the context can be done. Second, the tweets are compared on phrase, clause, sentence level to determine the similarity score. The tweet pair plays a role in deciding whether a topic is a rumor or not only if they have high semantic score.

The tweet pair with similar context is further processed by sentiment analysis. *Sentiment polarity* is calculated for every tweet in both the sets. A cumulative polarity of the tweets in the news tweet set is calculated as per the following formula:

$$newsPolarity = \begin{cases} positive & \text{if } p > n \\ negative & otherwise \end{cases}$$

where p = number of news channels tweets with postive polarity
n = number of news channels tweets with negative polarity

Then *mismatch ratio*, which reflects the degree of discrepancy between the news and public on that topic, is calculated as per the following formula:

$$mismatchRatio = \frac{M}{T}$$

where M = number of public tweets with polarity opposite to *newsPolarity*
T = Total number of public tweets

If the *mismatch ratio* is greater than a threshold value ρ(say 25%), the topic is labelled as a *rumor*. If a topic is being labelled as rumor, it means that public believes in the information that contrasts with the information from verified news channels and is thus posting/ retweeting it.

IV. IMPLEMENTATION AND RESULTS

A. Packages and Technologies Used

The algorithm was implemented in python programming language with help of other APIs and packages. We harvested trends and collect the tweet data by the Tweepy[2] wrapper in Python. The data collection has been implemented primarily by Tweepy and igraph[3], has been used for topic detection. NLTK[4] and Gensim[5] have been employed for semantic analysis and TextBlob package was used for sentiment mining.

B. Results

We implemented the proposed algorithm and were able to gather some rumors. One of the rumors, we found, was *Beef was being served in kerala house.* We manually checked it and it was a verified rumor(refer to Fig. 4).

Fig. 4: Verified news channel @EconomicsTimes confirming that *Beef was being served in kerala house.* is a rumor.

Among other rumors was, *Changing your facebook display picture to support Digital India is supporting internet.org*. This was also verified as a rumor(refer to Fig. 5).

Fig. 5: Verified news channel @abpnewstv confirming that *Changing your facebook display picture to support Digital India is supporting internet.org* is a rumor.

[2]www.tweepy.org
[3]www.igraph.org/python/
[4]www.nltk.org
[5]https://radimrehurek.com/gensim/

Also, our algorithm was able to classify tweets as credible news correctly, in cases when the news source polarity matched the general tweets about the topic, or when the mismatch ratio was lower than the threshold. For example, service tax to be increased to 14% form 15th November 2015 in India was certified by news sources and labelled as a credible topic. Also the High Court slashing 60% quota for top civil servants' kids in Sanskriti School was also found to not be a rumor.

Another rumor found was regarding the recent unfortunate Paris attacks, where a picture of a Sikh from Canada was photoshopped to show him holding a Quran and a suicide vest. This was also verified as a rumor as it mismatched with the news tweet set (refer to Fig. 6).

Fig. 6: Verified news channel @HuffpostUK confirms that *Paris attacker is a Sikh boy in the picture* is a rumor.

C. The Twitter Grapevine: Web Application

A web- application The Twitter Grapevine was developed to showcaseS the results of the algorithm and for providing an interactive portal to the twitter users to contribute to the rumor detection process. The application showcases multiple findings related to the rumor topic. Along with this functionality, it also allows users to give their view on whether the topic is a rumor or not. This interactivity helps in obtaining the views of users in real time and practically includes them in the labeling procedure, resulting in better classification, in addition to the algorithm labeling.

The application provides the following features to monitor the detected rumors and take informed decisions to take actions to curb its impact:

1) Sign in with Twitter: The application allows users to sign in using their Twitter account, so that they can follow users involved in the rumor and retweet their tweets, if they like.

2) Report as incorrect: The users can report the information on application as incorrect. The report is recorded and information is updated after manual inspection. This feature helps earn trust of the users by including them in the process. Also it enhances the accuracy of helps in enhancing the accuracy of the information on application.

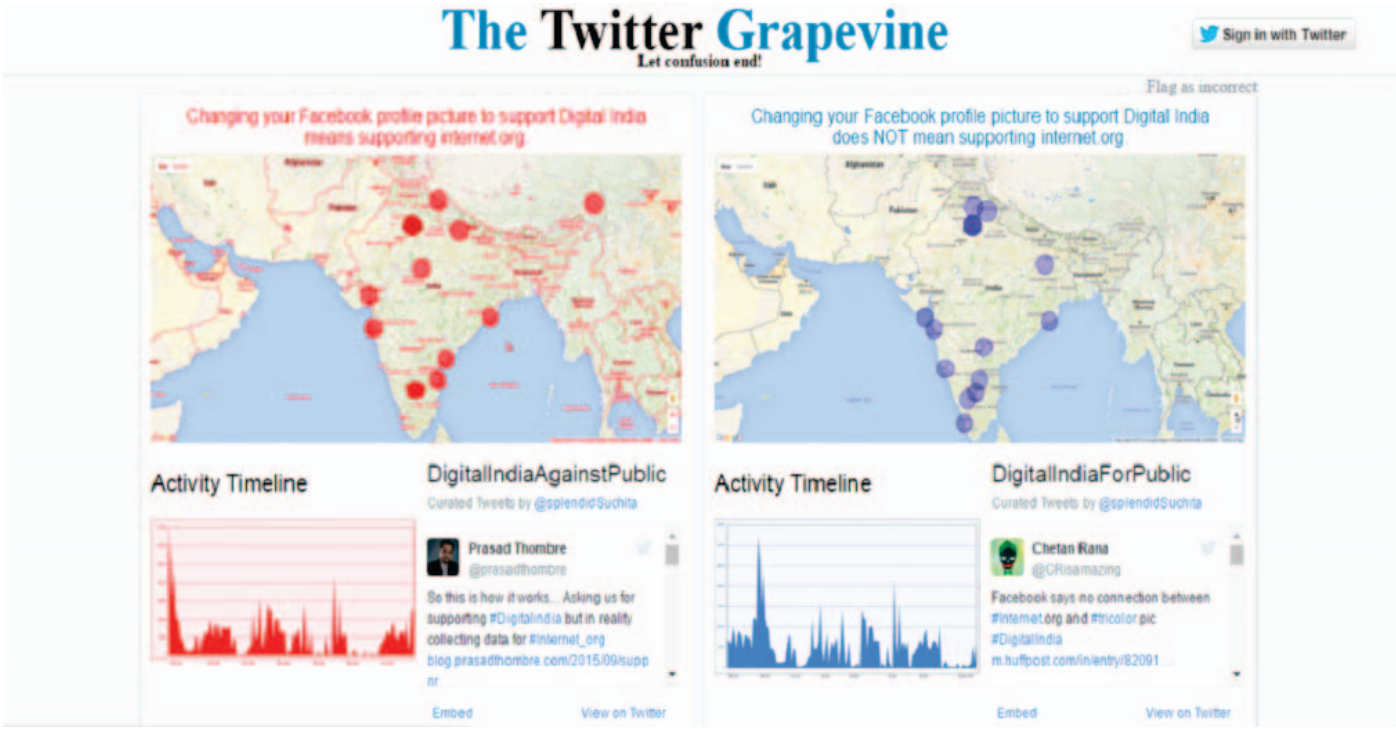

Fig. 7: The Twitter Grapevine, a screenshot of the application

3) Maps: The tweets locations are represented on a map, where two different maps are depicting the users sentiment and activity against or for the unverified information topic. The intensity of the markers represent the number of tweets being obtained from a location. The darker the marker, the more the number of tweets from that area.

4) Activity Timelines: The activity timelines for both the views are depicted on the topic page. These show the number of tweets on the topic, with a view as the time progresses.

5) Tweet Feeds: The application collects and displays all the tweets that have been found regarding the topic in consideration. The FOR tweets and the AGAINST tweets are segregated and displayed for the user to make an informed decision. The users can click and open up the tweets on Twitter and retweet the tweets which they want to propagate in the network. Thus users can actively participate in spreading the information they feel is right in the network.

V. ANALYSIS FOR ACCURACY

A. Changing your Facebook display picture to support Digital India is equal to supporting internet.org

		Predicted Label	
		Favourable	Unfavourable
Actual Label	Favourable	495	52
	Unfavourable	191	318

TABLE I: Accuracy analysis for Digital India and Facebook.org Rumor Topic

$$Accuracy = \frac{495 + 318}{1056} = 76.99\%$$

The results are better in this set of tweets since tweets extracted are more objective and less subjective. Thus the number of tweets with both semantic similarity and objectivity increases, increasing the number of correct classifications.

B. Beef is not being served in Kerala House

		Predicted Label	
		Favourable	Unfavourable
Actual Label	Favourable	1002	249
	Unfavourable	764	568

TABLE II: Accuracy analysis for KeralaHouse and Beef Rumor Topic

$$Accuracy = \frac{1002 + 568}{2583} = 60.78\%$$

The accuracy is low as compared to Digital India Rumor due to most of the tweets being subjective and very less number of tweets showing objectivity. Thus, a larger number of tweets are categorized as False Positives/ Negatives due to sentiment score above the range of +0.2 to -0.2.

VI. CONCLUSION

Through our work, we proved that it is possible to detect rumors on Twitter using tweets from the verified news channels as base.

A novel method to collect and harness huge Twitter data to detect rumors is presented. We then proposed approach to detect rumors from that data using sentiment and semantic analysis. We believe, on the basis of the results obtained that the proposed algorithm can detect rumors.

A working prototype is also presented to show the application of the proposed approach.

We believe that the proposed algorithm, if used to detect rumors especially in the critical times of emergency, can prove to be very useful to monitor and hence take actions to curb the spread of unverified information.

VII. OPEN ISSUES AND FUTURE WORK

A number of extensions are possible for our work. Attempts can be made to recover the Twitter network from rumor, thus trying to minimize the effect of rumor. The obtained accuracy can be improved by working on the semantic and sentiment analysis algorithms used. The problem is being recognised widely on not only Twitter, but almost all major social networks. The proposed approach can be extended to other social media after some modifications.

We plan to work in the above mentioned directions in future.

REFERENCES

[1] Mendoza, Marcelo, Barbara Poblete, and Carlos Castillo. *Twitter Under Crisis: Can we trust what we RT?*. Proceedings of the first workshop on social media analytics. ACM, 2010.

[2] Starbird, Kate, et al. "Rumors, false flags, and digital vigilantes: Misinformation on twitter after the 2013 boston marathon bombing." Conference 2014 Proceedings (2014).

[3] Dong, Wenxiang, Wenyi Zhang, and Chee Wei Tan. "Rooting out the rumor culprit from suspects." Information Theory Proceedings (ISIT), 2013 IEEE international Symposium on. IEEE, 2013.

[4] Vosoughi, Soroush. Automatic detection and verification of rumors on Twitter. Diss. Massachusetts Institute of Technology, 2015.

[5] Zhao, Zhe, Paul Resnick, and Qiaozhu Mei. "Enquiring Minds: Early Detection of Rumors in Social Media from Enquiry Posts." Proceedings of the 24th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2015.

[6] Zhou, Xing, et al. *Real-Time News Cer tification System on Sina Weibo* Proceedings of the 24th International Conference on World Wide Web Companion. International World Wide Web Conferences Steering Committee, 2015.

[7] Dayani, Raveena, et al. "Rumor: Detecting Misinformation in Twitter."