Topical Categorization of Credible Microblog Content

Raghul Vignesh Kumar A[1] Imthyaz Sheriff C[2]
[1,2]Department of Computer Science & Engineering
[1]AVS Engineering college, Salem, India [2]B.S.Abdur Rahman Univeristy, Chennai, India

Abstract— Microblogs are often considered to be useful sources of information. Increasing popularity of micro blogging sites like twitter, is making them a viable source of real-time news with a huge potential to replace news sites. Credibility of such microblogs is of concern when they are to be considered as sources of news. This paper details on an approach to examine microblogs like tweets, determine their credibility and categorize them under different news topics. This approach builds upon the previous work done in microblog credibility determination and rating. This paper works with twitter as test data source. The proposed novel approach can be extended to any other microblogs also.

Key words: Credibility, MicroBlogs, News Media

I. INTRODUCTION

Rapid evolution and prevalence of online social networking and micro-blogging mediums has resulted in making the internet a viable competition to traditional media like television and print media as a source for obtaining news and information about current events. Share information and expressing opinions on the fly through online social networks and microblogs have become the order of the day. Not only the people use the giant micro-blog services, but news agencies and corporations have also started to use it to disseminate news and advertisement.

The growth of microblogging social networks like Twitter has led to a huge number of generated contents (tweets) every second. A study [5] on a real-world dataset reveals that about 26% of the trending topics rising from Twitter "as-is" are also found as hot queries issued to Google, 72% of the similar trends appear first on Twitter. Twitter has can be used as a sensor to gather up-to-date information about the state of the world. For example, Sakaki *et al.* ,[4] used Twitter for early detection of earthquakes in hope of sending alert about them before it becomes a catastrophe. In 2013, Twitter [8] attracted over 200 million users whom publish more than a billion posts every week. In a study[1] author has suggested that twitter covers most of the events that are reported by newswire providers and that many events reported in Twitter are not mentioned in newswire. According to [6] Twitter represents a disruptive technology with regard to primary news that is to the detriment of newspaper companies. This paper approaches microblogging sites like twitter with a novel idea of making them into a credible source of news. Similar to topic wise categorization of new items in news media sites, this paper intends to categorize microblog content under different topics after determining their credibility.

Microblog information often spreads faster and reaches out to a wider audience due to its broadcast nature when compared to traditional news media. This is further accelerated by ever increasing use of smartphones and tablets. Microblogger can provide real-time information (via smartphones and tablets) from the actual location where the real life events of interest are unfolding. People all across the world are increasingly using micro blogging sites like Twitter to share opinions, news, advice, moods, facts, rumors, current affairs and everything else imaginable. One key difference between dissemination of information or news through traditional media and microblogging services like Twitter is that, it is a crowd-sourced medium which presents the challenge of determination of credibility.

In case of television or print or news websites the source of information are few and known (i.e. credible). On the contrary, users on microblogging sites are numerous, they act like its sensors which keep collecting nuggets of information, which may or may not be credible. Trustworthiness of content generated on microblogging sites like Twitter is questionable due to the unmonitored, anonymous nature of the Internet. This is a serious problem, which when unchecked can result in spreading of fake news, false alarms and hoax in this connected world, where microblogs are acting as news sources for more and more people every day. Many journalists have hailed the immediacy of the microblogging services which allowed "to report breaking news more rapidly than most mainstream media outlets". Thus, determination of content credibility in microblogs is of integral importance to any approach which aims to present microblog content as a source on news.

II. RELATED WORK

Cha et. al. [7] found that Twitter brings a playing field together for all three voices: the mass media(BBC etc), evangelists (Celebrities, politicians)and grassroots (Ordinary users). On one hand, the mass media play a dominant role in the network. They excel at all aspects of news spreading; they have many followers, their links are well reciprocated, and they have topological advantages to collect diverse opinion of other users. Their tweets also reach a large portion of the audience directly, without the involvement of other influential users. On the other hand, the mass media in Twitter, unlike the traditional media networks, are not necessarily the first to report events. In some cases, in fact, it is the small, less connected grassroots or evangelists that trigger the spreading of news or gossip, even without the mass media's coverage of such topic. Evangelists, overall, played a leading role in the spread of news in terms of the contribution of the number of messages and in bridging grassroots who otherwise are not connected.

Petrovi et. al. [1] found that Twitter appears to cover nearly all newswire events, but newswire only covers a subset of the events reported in Twitter. Most of the tweeted events are sports-related, have value only for a short time or to a very restricted audience, and would thus lose value by the time they are reported in the newswire. For major events neither stream consistently leads the other in terms of reporting time, i.e. both sources can report news first. Meanwhile, for the automatically detected events, there are very few where Twitter leads newswire when both

sources report a story. Twitter reported before newswire into seven broad categories, namely: politics, sports, disasters & accidents, business & economy, entertainment, technology, and other.

Giummolè et. al. [5] claimed that a trending topic on Twitter could later become a hot query on Google as well. Indeed, information flooding nearly real-time across the Twitter social network could anticipate the set of topics that users will be interested in – thereby will search for – in the near future. To validate this claim, following contributions provided. First, the Trend Bipartite Graph (TBG) to represent the lexical similarity between any pair of social and web trends, as extracted from real-world. Twitter and Google datasets. The TBG used a threshold to link those trends that were most likely related. Then the ability of Twitter in predicting and causing a Google trend was measured by conducting an exhaustive comparison of several time series regression models.

Raghul et. al [2] proposed a comprehensive framework for credibility determination of microblogs by integrating several diverse. This framework follows a modular approach where several credibility determination methodologies can be deployed either in a standalone manner or applied integratively. Modules based on this framework are user feature extraction, content feature extraction, topic classifier,RSS feed correlator to determine the credibility of microblogs and to be delivered with relevant form of ratings and other cues like accept\reject visual tags to the analyzed tweets.

Bellaachia et. al. [14] focused on text sparseness. Due to the sparseness in text and short length of tweets, it is always limited without enough knowledge contexts in getting the information. Bellaachia et. al. proposed to augment each topical graph by using the ubiquitous links appearing in tweets as hashtags. By using an auxiliary set of tweets it is proved that performance can be improved. It involved using of two different methods for selecting hashtags, first, a simple method by using the frequency of hashtags appearing in tweets is suggested. Second, a more sophisticated approach using document similarity techniques and measuring the relevancy of each tweet to the target set is proposed with significant improvements.

Han et. al. [10] analyzed the news access/sharing and related query behavior on Twitter and the usual Web. It was found that the highly accessed news is not necessarily shared on Twitter. Twitter users tend to share the politics, IT and economics/business news and the fluctuation of category of often shared news is more visible.

III. HASHTAGS

A hashtag is defined to be a word or phrase prefixed with the symbol "#". It is widely used in current social media sites including Twitter, Google+, and new feeds as a significant metatag to categorize\group users' messages, to propagate ideas and topic trends. The use of hashtags has become an integral part of the social media culture. Searching Twitter for a hashtag returns all the recent tweets that include it. Given all this, and given that hashtags are now often seen "in the wild" (on other social networks, such as Facebook, and even off-line on T-shirts), the hashtags are gaining a strong technically, culturally, and linguistically.

Weng et. al. [11] focuses on measuring the interestingness of hashtags in Twitter. Twitter allows users to assign hashtags to tweets so that they can be navigated and searched more easily. Due to large number and variety of new tags the get created each day, not all of them tend to be meaningful and interesting for serious consumption. The ability to filter out a set of interesting tags can lead to several useful applications. Weng et. al. proposed MEDIC, which measures using hashtags' interestingness by studying how they are discussed within and across communities. Experimental studies showed that MEDIC achieves a fairly good performance with still a good space for improvement.

Ma et. al. [12] proposed a topic model, TLDA(Tag-Latent Dirichlet Allocation) to address these two challenges: first, how to interpret hashtags and second how to find related hashtags. The TLDA model learns the hidden topic structures for each hashtag and measure the similarities between every pair of hashtags, as hastags are modeled in a common topic space. TLDA was applied to extract meaningful topics from tweets, and use visualization techniques to understand hashtags. It also show that this method can help discover a group of hashtags created to describe one common event based on the topic similarity of the hashtags.

Bellaachia et. al. [13] used an extensive preprocessing approach to and introduced new features that might improve topical keyphrase extraction in Twitter by leveraging hashtags. Consequently, Bellaachia et. al. proposed a novel unsupervised graph-based keyword ranking method that considers words weight in addition to edges when calculating the ranking. The potential and validity of both approaches have been demonstrated by conducting an experimental evaluation which produced an improvement when applying NE-Rank and hashtags enhanced extraction.

IV. ARCHITECTURE

This paper provides an integrated approach for credibility determination, followed by topical categorization of microblog contents like tweets. This framework is illustrated in figure 1 shown below.

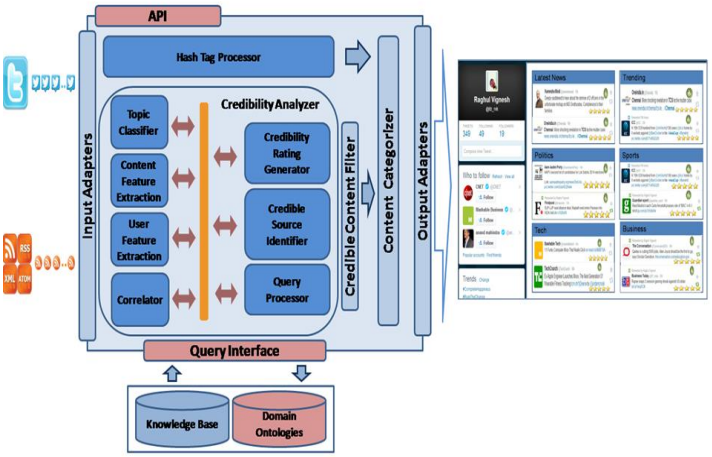

Fig. 1: Credible Microblog Newswire Framework

This framework follows a modular approach where several credibility determination methodologies can be deployed in generating credible newswire alternative to existing online news feeds. This approach makes use of content features, user features for creditability determination of tweets in a standalone manner and makes use of RSS feeds for further correlative credibility analysis and assertion with news sites. This approach also makes use of Hash Tags for topical classification using Tag Latent Dirichlet Allocation.

A. Credibility Analyzer

This is the core component which incorporates the integration logic for comprehensive credibility analysis. It includes individual functional modules like topic classifier, content feature extraction, user feature extraction, correlator, credibility rating generator, credible source identifier and Query processor. It makes use of input adapters for subscriptive access to tweets and feeds (RSS, ATOM, etc) . Output adapters are used to generate suitable UI deliverables. Core functionality of this analyzer is to determine the credibility of the content and give a rating as to how credible it is.

Analysis of microblog credibility is based on user features, content features and correlation with RSS feeds from standard and popular news sites. Raghul et. al. [2] provides the detailed working the credibility analyzer. At a very simplest level of abstraction, credibility analyzer gives rating to the content (tweets) based on the credibility score arrived. Credibility analyzer also forms the basis for determining credible sources of news pertaining to a particular category. Credibility analyzer also facilitates visual tagging of tweets based on whether they are credible or not credible. Sample tweets after credibility analysis are indicated in figure 2 and figure 3.

B. Credibility Content Filter

Apparently not all credible tweets should be presented due to the redundancy problems, so credible content filter evaluates the content (tweets) based on the maximum credibility score, which results in filtering out the best, most credible tweets from the entire credible tweet collection. For example, the user can set the threshold level as 3 and above (the actual scale of rating is 1 to 5) for his or her profile. Only credible tweets having a rating of 3 or above will be chosen for further processing by the Content Categorizer. Figure 2 shows a sample tweet with a credibility rating of '1'. This will be filtered out i.e reject by content filter and will not be listed in any category.

Fig. 2: Sample Tweet with Visual Cue for a Not Credible Tweet

Figure 3 shows a sample tweet with a credibility rating of '5'. This will be filtered in i.e accepted by content filter and will get listed in appropriate category.

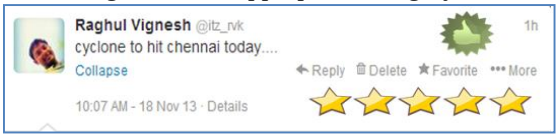

Fig. 3: Sample Tweet with Visual Cue for a Credible Tweet

C. Knowledgebase and Domain Ontologies

It consists of strategies to be followed to determine the credible content based on past knowledge. Framework significantly depends on knowledge base spanning several domains to determine the credibility. Domain specific, credible RSS\ATOM feed sources are available in the knowledge base. This enables correlation analyzer effective in making news subscriptions and also in identifying relevant feeds for correlation.

Ontology formally represents knowledge as a set of concepts within a domain, using a shared vocabulary to denote the types, properties and interrelationships of those concepts. Selected, popular domain ontology are made available. These are used to segregate the tweets based on the topic of the tweet and also in enabling relevant correlation with feeds.

D. Hash Tag Processor

Hash tag processor extracts hash tags from tweets and identifies topics based on topical modeling using Tag-LDA. These identified topics will be used for grouping of filtered, credible tweets. Topic categories will dynamically change over a period of time based on the nature of incoming tweets. For better topical correlation, hashtags from RSS/ATOM feeds can also be extracted and analyzed with relevant tweets

E. Content Categorizer

After filtering credible tweets which meet a certain threshold it is further classified based on the content topic and hashtags of tweet with the help of hash tag processor. Each of the tweets are listed in different columns such as Latest news, trending, politics, sports, business, technology etc. with the help of knowledge base and domain ontologies. Tweets which are high in frequency are listed under 'trending'; where as recent tweets will get listed under 'latest' after credibility determination and filtering. Number of tweets under each category will vary and depend on the threshold set for the content filter by the user via his profile settings.

F. Input Adapters

Functionality of input adapters is to subscribe and listen to incoming streams like tweets, RSS and convert them into desired format to be processed by other modules of credible determination framework. Input adapter deal with initial preprocessing and suitable format conversion. They also support the correlation analyzer by providing on the fly subscription to desired RSS\ATOM feeds.

G. Ouptut Adapters

Results of various modules of Credible Microblog Newswire Framework have to be delivered in a user interface with relevant form of ratings and accept\reject visual tags labeled under corresponding topics as shown in the figure 4.

V. Summary And Future Work

This paper proposed an integrated approach for topical categorization of credible microblog content. This approach extend the previous work done for credibility determination by making use of hash tags and Tag Latent Dirichlet Allocation. This resulted in a newswire like system capable of turning microblogs into sources of credible news.This

approach can be extended and customized for other text based microblogs.

Future work involves adopting Complex Event Processing(CEP) paradigm for instantaneous determination of credible news content from microblogs. CEP can subscribe to multiple microblogging sites and extract content on the fly. It can also hook on the multiple RSS/ATOM feeds on a need basis and provide analytical processing. Such approach will increase the overall efficiency and trustworthiness of this system.

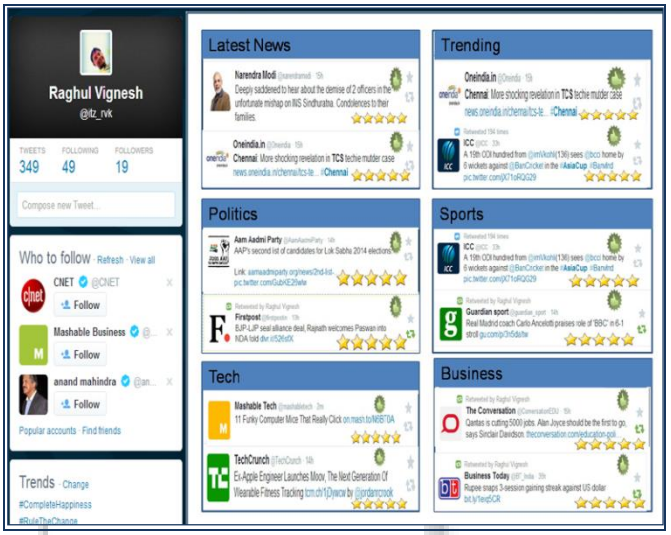

Fig. 4: Categorized Credible Mircoblog News - Sample View

REFERENCES

[1] Sasa Petrovi, Miles Osborne,Richard McCreadie,Craig Macdonald, Iadh Ounis,Luke Shrimpton,"Can Twitter replace Newswire for breaking news?", Association for the Advancement of Artificial Intelligence, www.aaai.org,2013.

[2] Raghul Vignesh Kumar A and Imthyaz Sheriff C "Determining Content Credibility in Micro Blogs",1st International Conference on Business Analytics and Intelligence, IIMB, Bangalore, December 11-13, 2013.

[3] Lin, R. Snow, and W. Morgan, "Smoothing techniques for adaptive online language models: topic tracking in tweet streams," in Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pp.422–429,2011.

[4] Sakaki, T. Okazaki, M. and Matsuo, Y. "Earthquake shakes Twitter users: real-time event detection by social sensors". In Proceedings of the 19th international conference on World wide web, ACM, pp.851-860, 2010.

[5] Federica Giummolè, Salvatore Orlando, Gabriele Tolomei,"Trending Topics on Twitter Improve the Prediction of Google Hot Queries"978-0-7695-5137-1/13 2013,IEEE, 2013.

[6] Yasuyoshi Aosaki, Taro Sugihara, Katsuhiro Umemoto,"Examining the Trend toward a Service Economy in Information Media through Changes to Technology: Influence of Twitter on Media Companies", 978-1-890843-21-0/10,IEEE,2010.

[7] Meeyoung Cha, Fabrício Benevenuto,Hamed Haddadi, And Krishna Gummadi,"The World Of Connections And Information Flow In Twitter",IEEE transactions on systems, man, and cybernetics—part a: systems and humans, vol. 42, no. 4, July 2012.

[8] http://en.wikipedia.org/wiki/Twitter

[9] M. Hu, S. Liu, F. Wei, Y. Wu, J. Stasko, and K.-L. Ma. "Breaking news on Twitter". In The Proceedings of Annual Conference on Human Factors in Computing Systems, pp .2751–2754,ACM, 2012.

[10] Hao Han, Hidekazu Nakawatase and Keizo Oyama,"An Exploratory Analysis of Browsing Behavior of Web News on Twitter". 978-0-7695-4938-5/12 2012 IEEE,2012.

[11] Jianshu weng, Ee Peng Lim, Qi He, and Cane wing-Ki Leung ,"What Do PeopleWant in Microblogs? Measuring Interestingness of Hashtags in Twitter" International conference on datamining,IEEE,2010

[12] Zhiqiang Ma, Wenwen Dou, Xiaoyu Wang, Srinivas Akella ,"Tag-Latent Dirichlet Allocation: Understanding Hashtags and Their Relationships", International Conferences on web Intelligence (Wi) and intelligent Agent Technology (IAI') ,IEEE/WIC/ACM, 2013.

[13] Abdeighani Bellaachia and Mohammed Al-Dhelaan "NE-Rank: A Novel Graph-based Keyphrase Extraction in Twitter", International Conferences on web Intelligence and Intelligent Agent Technology, IEEE/WIC/ACM,2012.

[14] Abdelghani Bellaachia and Mohammed Al-Dhelaan ,"Learning from Twitter Hashtags: Leveraging Proximate Tags to Enhance Graph-based Keyphrase Extraction", IEEE International Conference on Green Computing and Communications, Conference on Internet of Things, and Conferenceon Cyber, Physical and Social Computing,IEEE,2012

