International Journal of Engineering Science Invention (IJESI)
ISSN (Online): 2319 – 6734, ISSN (Print): 2319 – 6726
www.ijesi.org || PP. 46-51

Topic Detection on Social Media Posts

Sandhiya.R[1] and K..Manohari [2]
[1](*Department of Computer Science, TheivanaiAmmal College for Women, India)*
[2](*Department of Computer Science, TheivanaiAmmal College for Women, India)*

Abstract: *The widespread of the Internet revolution, the discovery of valuable information from online resources has become prominent research area. Generally, the knowledge derived from the social events, in specific to, social news media, which are reliable according to its content. On the other hand, the Internet, being a free and open forum for information exchange, has recently seen an extremely interesting phenomenon known as social media. In social media, regular, non-journalist users are able to publish unverified content and express their interest in certain events. This study focuses on recognizing relevant topic to the news event posted in Twitter. At first, the generated social news are collected and preprocessed to remove the irrelevant tags and noises. The considered social factors are Media Focus (MF), User Attention (UA) and User Interaction (UI) is processed over the collected twitter events. Media Focus depicts the prevalence of the discussed topics, User Attention (UA) depicts the temporal prevalence of the current topics and User Interaction (UI) discusses the strength of the current topics. These factors are analyzed and it's ranked under unsupervised frameworks. Finally, this article has achieved a consolidated, filtered and ranked social media events occurring in the Twitter dataset.*

Keywords: *Twitter, social media, social news, Social factors, Topic discovery and unsupervised frameworks*.

I. Introduction

The Internet has subverted the autocratic way of disseminating news by traditional media like newspapers. Online trends are different from traditional media as a method for information propagation. For instance, Google Hot Trends ranks the hottest searches that have recently experienced a sudden surge in popularity. Meanwhile, these trends may attract much more attention than before due to their appearance on Google Hot Trends. Microbloggings have become one of the most popular social media outlets. One micro blogging service in particular, Twitter, is used by millions of people around the world, pro-viding enormous amounts of user-generated data. One may assume that this source potentially contains information with equal or greater value than the news media, but one must also assume that because of the unverified nature of the source, much of this content is useless. For social media data to be of any use for topic identification, we must find a way to filter uninformative information and capture only information which, based on its content similarity to the news media, may be considered useful or valuable.

The temporal prevalence of a particular topic in the news media indicates that it is widely covered by news media sources, making it an important factor when estimating topical relevance. This factor may be referred to as the MF of the topic. The temporal prevalence of the topic in social media, specifically in Twitter, indicates that users are interested in the topic and can provide a basis for the estimation of its popularity. This factor is regarded as the UA of the topic. Likewise, the number of users discussing a topic and the interaction between them also gives insight into topical importance, referred to as the UI. By combining these three factors, we gain insight into topical importance and are then able to rank the news topics accordingly.

Previous research has studied trend taxonomy, trend detection and real events extraction from Twitter trends. However, researchers have paid little attention to twitter trend manipulation.It is reported that attackers manipulate Google trends by simply employing large group of people to visit Google and search for a specific keyword phrase.Thus, the focus of this work is on Twitter trend manipulation. In this paper, the primary questions we attempt to answer are whether the malicious users can manipulate the Twitter trends and how they might be able to do that? Being exposed to real-time trending topics, users are entitled to have insight into how those trends actually go trending. Moreover, this research also cast light on how to enhance a commercial promotion campaign by reasonably using Twitter trends. To investigate the possibility of manipulating Twitter trends, we need to deeply understand how trending works twitter.

Twitter states that trends are determined by an algorithm and are always topics that are immediately popular. However, the detailed trending algorithm of Twitter is unknown to the public, and we have no way to find out what it specifically is. Instead, we study Twitter trending at the topic level and infer the key factors that can determine whether a topic trends from its popularity, coverage, transmission, potential coverage, and reputation. After identifying those key factors that are associated with the trends, we then investigate the manipulation and countermeasures from the perspective of these key factors. The major contributions of this

work are as follows: The evidence of the existing manipulation of Twitter trends. In particular, employing an influence model, we analyze the dynamics of an endogenous hash tag and identify the manipulation from its endogenous diffusion. After further investigating the manipulation in the dynamics, we disclose the existence of a suspect spamming infrastructure.

The corresponding dynamics for each factor above are extracted, and then Support Vector Machine (SVM) classifier is used to check how accurately a factor could predict trending. In propose an unsupervised system—SociRank—which effectively identifies news topics that are prevalent in both social media and the news media, and then ranks them by relevance using their degrees of MF, UA, and UI.Even though this paper focuses on news topics, it can be easily adapted to a wide variety of fields, from science and technology to culture and sports. To the best of our knowledge, no other work attempts to employ the use of either the social media interests of users or their social relationships to aid in the ranking of topics. Moreover, SociRank undergoes an empirical framework, comprising and integrating several techniques, such as keyword extraction, measures of similarity, graph clustering, and social network analysis. The effectiveness of our system is validated by extensive controlled and uncontrolled experiments.

II. SVM Classifiers

The corresponding dynamics for each factor above are extracted, and then Support Vector Machine (SVM) classifier is used to check how accurately a factor could predict trending. We find that, except for transmission, each studied factor is associated with trending. We further illustrate the interaction pattern between malicious accounts and authenticated accounts, with respect to trending. The Paper present the threat of malicious manipulation of Twitter trending, given compromised and fake accounts in the suspect spamming infrastructure we observed. Then we demonstrate how compromised and fake accounts could threaten Twitter trending by simulating the manipulation of dynamics as compromised and fake accounts would do.

III. Dataset

3.1 Data Collection

Data collected our dataset via Twitter API through two different collection windows. One lasted for 40 days and the other lasted for 30 days. At the end, we obtained more than 69 million tweets from 5 million accounts. Since we focus on the hash tags, we only analyze the tweets with hash tags. More specifically, our dataset was collected via Stream API. We also collected the public trends of Twitter via Rest API. Sample Stream and Search Stream. We obtain a sample stream via Twitter's Streaming API. We define the 15 most frequent hash tags in the sample stream as sample trends. Sample trends are retrieved from the sample stream every 30 minutes. We create a search stream by opening up a new streaming channel via Streaming API and searching sample trends. Therefore, the sample stream and search stream are not inclusive of each other, since they are from two different streaming channels of the Streaming API.

Public Trends and Sample Trends. Twitter trends include trending hash tags and trending keywords. Our focus is on the trending hash tags. Thus, the trends in the rest of the paper represent trending hash tags only. Public trends are published by Twitter and available via the Twitter API. Sample trends are obtained by ranking the frequency of hash tags over the sample stream.Note that, throughout this paper, trends represent public trends if not specified. The trends used to conduct trending analysis are the intersection of sample trends and public trends. Sample Dynamics and Search Dynamics. We define the dynamics of a topic as the variation of the topic against time with respect to a specific frequency feature, such as tweet number or account number. For a certain topic, we obtain its dynamics through its sample stream and search stream independently. Sample dynamics represent how the topic evolves in the sample stream, while search dynamics reflect the evolution of the topic in the search stream.

3.2 Validation of Dataset

The major objective of this work is to study the key factors of Twitter trending and inspect the possible manipulation of these factors. In this respect, we validate the representativeness of our dataset in two ways. On one hand, sample trends are supposed to reflect the public trends to a certain extent; on the other hand, the syncretization of sample dynamics and search dynamics should be able to embody the critical information for inferring the key factors of Twitter trending.

3.3 Endogenous Factors and Manipulation

The employ, an influence model (Linear Influence Model, LIM) to capture the network effect on the spread of the memes. LIM is used to model the global influence of a node (an account) on the rate of diffusion through a network, which can be expressed as

$$V (t + 1) =Xu2A(t)Iu(t \quad tu); (3)$$

where V (t + 1) represents the number of nodes that are influenced at time t + 1, A(t) denotes the set of nodes that have already been influenced before time t, and I u (l) is the influence function of node u at lth time step after it is influenced at time t u (tu< t). LIM has been evaluated that, for the memes mentioned above, most of the observed dynamics could be attributed to the influence of nodes, especially considering the imitation factor b(t):

$$V (t + 1) =Xu2A(t)Iu(t \quad tu) + b(t): (4)$$

The imitation means that nodes imitate one another because the topic is popular and everyone talks about it. However, for the memes, the imitation happens only due to the spread in the network. Therefore, we exclude imitation from the model and take the manipulation ex(t) into account. The influence model we consider is

$$V (t + 1) =Xu2A(t)Iu(t \quad tu) + ex(t): (5)$$

Extensive research has been done on the influence in Twitter. Researchers not only inspected the effectiveness of different influence measures, such as follower number, tweet number, and mention number, but also proposed algorithms to measure the influence.

IV. Proposed Technique

In the Article propose an unsupervised system—SociRank—which effectively identifies news topics that are prevalent in both social media and the news media, and then ranks them by relevance using their degrees of MF, UA, and UI. Even though this paper focuses on news topics, it can be easily adapted to a wide variety of fields, from science and technology to culture and sports. To achieve its goal, SociRank uses keywords from news media sources (for a specified period of time) to identify the overlap with social media from that same period. We then build a graph whose nodes represent these keywords and whose edges depict their co-occurrences in social media. The graph is then clustered to clearly identify distinct topics. After obtaining well-separated topic clusters (TCs), the factors that signify their importance are calculated: MF, UA, and UI. Finally, the topics are ranked by an overall measure that combines these three factors.The major contributions of this work are as follows: We demonstrate the evidence of the existing manipulation of Twitter trends. In particular, employing an influence model, we analyze the dynamics of an endogenous hash tag and identify the manipulation from its endogenous diffusion. After further investigating the manipulation in the dynamics, we disclose the existence of a suspect spamming infrastructure.

In this article Twitter trending at topic level, considering topics' popularity, coverage, transmission, potential coverage, and reputation. The corresponding dynamics for each factor above are extracted, and then Support Vector Machine (SVM) classifier is used to check how accurately a factor could predict trending. We find that, except for transmission, each studied factor is associated with trending. We further illustrate the interaction pattern between malicious accounts and authenticated accounts, with respect to trending.The Article present the threat of malicious manipulation of Twitter trending, given compromised and fake accounts in the suspect spamming infrastructure we observed. Then we demonstrate how compromised and fake accounts could threaten Twitter trending by simulating the manipulation of dynamics as compromised and fake accounts would do.

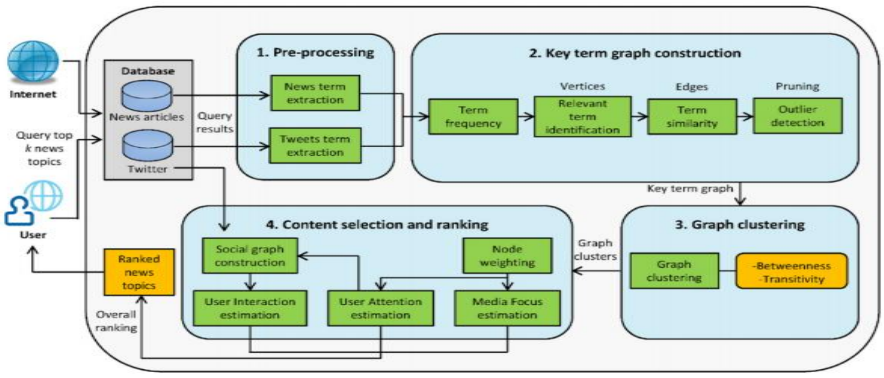

Fig: Framework Diagram

IV. Experimental Analysis

The testing dataset consists of tweets crawled from Twitter public timeline and news articles crawled from popular news websites during the period between November 1, 2013 and February 28, 2014. The news

websites crawled werecnn.com, bbc.com, cbsnews.com, reuters.com, abcnews.com, and usatoday.com. Over the specified period of time, a total of 105 856 news articles and 175 044 074 bilingual tweets were collected. After non-English tweets were discarded, 71 731 730 tweets remained.

The dataset was divided into two partitions.
1) Data from January and February 2014 were used as the testing dataset, on which experiments were performed for the overall method evaluation.
2) Data from November and December 2013 were used as the control dataset, where experiments were performed to establish adequate thresholds and select measures that presented the best results.

TABLE I
SOME STATISTICS RELEVANT TO THE TESTING DATASET

Time period	# topics	Avg. tweets	Avg. news	Avg. users
2014/01/01–10	84	2138	17	430
2014/01/11–20	112	1585	13	788
2014/01/21–30	100	2615	20	1626
2014/02/01–10	99	3113	17	1190
2014/02/11–20	106	3567	12	932
2014/02/21–28	79	2386	16	398
Average	97	2567	16	894

V. Related Works

Topic detection methods can be broadly classier into two categories: document-pivot methods, which detect topics based on semantic distance between two documents, and feature-pivot methods, which track words and discover topics by grouping words together [19], [16]. To the best of our knowledge, this is the first effort to investigate whether Twitter trends could be manipulated. Research on trending topics in Twitter includes real event reorganization real time trending topic detection the evolution of trending topic characterization and the taxonomy of trending topics.[1],[8]. Temporal and Social Term Evaluation is an event detection system based on keyword-lifecycle which automatically generates events from tweets [20]. Becker etal. Analyzed the stream of Twitter messages and distinguished the messages about real events from non-event messages based on a clustering method [6]. Zubiagaetal.Categorized different triggers that leverage the trending topics by using social features rather than content-based approaches [7]. In the detection of real time trending topics, Agarwaletal.identified the emerging events before they became trending topics by modeling the detection problem as discovering dense clusters in highly dynamic graphs Kasiviswanathanetal. Presented dictionary learning based framework for detecting emerging topics in social media via the user-generated stream [8]. J., Mohebbi Energy functions to model the life activity of news events on Twitter and proposed a news event detection method based on online energy function. Cataldietal identified emerging terms from user content by measuring user authority and proposing a keyword life cycle model, and then detected the emerging topics by formalizing the keyword-based topic graph [2]. To address the evolution and taxonomy of trending topics, Altshuler and Pan presented the lower bounds of the probability that emerging trends successfully spread through the scale-free networks is told by Nikolov [3]. As sure etal studied trending topics on Twitter and theoretically analyzed the formation, persistence, and decay of trends. Naamanetal. Characterized the trends in multiple dimensions and presented taxonomy of trends. They also proposed a collection of hypotheses on different kinds of trends and evaluated them [9]. Lehmann etal.classified the popular hash tags by the temporal dynamics of hash tags. Iranietal. Focused on the trend stuffing issue and developed a classifier to automatically identify the trend-stuffing in tweets. Whether a topic begins trending is closely related to the influence of users who are involved with the topic and (2) the topic adoption for users who are exposed to the topic. Cha et al. performed a comparison of three different measures of influence: in degree, tweet, and mention. Cataldi[15]. Proposed a topic-sensitive Page Rank measure for user influence. Romero et al. proposed an algorithm to measure the relative influence and passivity of each user from the viewpoint of a whole network. Bakshyetal.Measured the influence from the diffusion tree. The studies of topic adoption in Twitter Mainly concentrate on hash tag adoption. Lin et al. classified the adoption of hash tags into two classes and proposed a framework to capture the dynamics of hash tags based on their topicality, interactivity, diversity, and prominence [1]. Yang etal.Studied the effect of the dual role of a hash tag on hash tag adoption. While there has been prior work done on event detection on Twitter domain, most of the works either perform a post-hoc analysis of tweets, which detect events with significant time lag after they have happened [6].Twitter is an online social networking service and micro blogging service that enables its users to send and receive text based messages of up to 140 characters, known as tweets. Twitter has 280 million active monthly users, generating over 500 million tweets daily . Their large user base and the ease of use make it one of the fastest and most popular information sources [21].Keyword-lifecycle based events are also used to compare the recent behavior of a word to standard historic behavior of the word to identify abnormality. Aliello et al. presented an extensive survey of event detection models on data gathered from Twitter [22].This approach is computationally in efficient and is not scalable, because of the time taken to compute signals for each word and calculate

correlation. Tweet tokenizes the words using Microsoft N-gram service and Wikipedia to calculate importance of the word [23]. More recent approaches use probabilistic models to detect clusters of words to represent events; an example is Latent Dirichlet Allocation (LDA) [24]. A SVM model based on tweet length, frequency of a word and co-occurrence frequency is then used to classify if the word represents an event or not. Sakaki et al. proposed event detection as a classification problem [25].

VI. Conclusion

In this paper, we proposed an unsupervised method— SociRank—which identifies news topics prevalent in both social media and the news media, and then ranks them by taking into account their MF, UA, and UI as relevance factors. The temporal prevalence of a particular topic in the news media is considered the MF of a topic, which gives us insight into its mass media popularity. The temporal prevalence of the topic in social media, specifically Twitter, indicates user interest, and is considered its UA. Finally, the interaction between the social media users who mention the topic indicates the strength of the community discussing it, and is considered the UI. To the best of our knowledge, no other work has attempted to employ the use of either the interests of social media users or their social relationships to aid in the ranking of topics.With the datasets we collected via Twitter API, we first evidence the manipulation of Twitter trending and observe a suspect spamming infrastructure. Then, we employ the SVM classifier to explore how accurately five different factors at the topic level (popularity, coverage, transmission, potential coverage, and reputation) could predict the trending. We observe that, except for trans-mission, the other factors are all closely related to twitter trending. We further investigate the interacting patterns between authenticated accounts and malicious accounts.

Finally, we present the threat posed by compromised and fake accounts to Twitter trending and discuss the corresponding counter measures against trending manipulation. In the Paper, they performed extensive experiments to test the performance of SociRank, including controlled experiments for its different components. SociRank has been compared to media-focus-only ranking by utilizing results obtained from a manual voting method as the ground truth. In the voting method, 20 individuals were asked to rank topics from specified time periods based on their perceived importance. The evaluation provides evidence that our method is capable of effectively selecting prevalent news topics and ranking them based on the three previously mentioned measures of importance. Our results present a clear distinction between ranking topics by MF only and ranking them by including UA and UI. This distinction provides a basis for the importance of this paper, and clearly demonstrates the shortcomings of relying solely on the mass media for topic ranking.

References

[1]. Wall Street Journal (Inside a Twitter Robot Factory), http://online.wsj.com

[2]. Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., and Brilliant, L. Detecting influenza epidemics using search engine query data. Nature, 457(7232), 1012-4.

[3]. Nikolov, S. Trend or No Trend: A Novel Nonparametric Method for Classifying Time Series (Doctoral dissertation, Massachusetts Institute of Technology).

[4]. Just, M., Crigler, A., Metaxas, P., and Mustafaraj, E. It's Trending on Twitter-An Analysis of the Twitter Manipulations in the Massachusetts 2010 Special Senate Election. In APSA 2012 Annual Meeting Paper.

[5]. Ratkiewicz, J., Conover, M., and Miss, M. Detecting and tracking the spread of AstroTurf memes in microblog streams. 5th International Conference on Weblogs and Social Media, 2010.

[6]. Becker, H., Naaman, M., and Gravano, L. beyond trending topics: Real-world event identification on twitter. ICWSM 2011.

[7]. Zubiaga, A., Spina, D., and Martinez, R. Classifying Trending Topics: A Typology of Conversation Triggers on Twitter. CIKM 2011.

[8]. Agarwal, M. K., Ramamritham, K., and Bhide, M. Identifying Real World Events in Highly Dynamic Environments. VLDB 2012.

[9]. Naaman, M., Becker, H., and Gravano, L. Hip and trendy: Characterizing emerging trends on Twitter. Journal of the American Society for Information Science and Technology, 62(5), 902-918.

[10]. Lee, K., Palsetia, D., Narayanan, R., Patwary, M. M. A., Agrawal, A., and Choudhary, A. Twitter Trending Topic Classification. 2011 IEEE 11th International Conference on Data Mining Workshops,251-258.

[11]. Morstatter, F., Ave, S. M., and Carley, K. M., Is the Sample Good Enough? Comparing Data from Twitters Streaming API with Twitters Firehose, AAAI 2013.

[12]. Cover, T.M. and Thomas, J.A., Elements of information theory, John Wiley and Sons, 2012.

[13]. Kasiviswanathan, S. P., Melville, P., Banerjee, A., and Sindhwani, V. Emerging topic detection using dictionary learning. CIKM 2011.

[14]. Lu, R., Xu, Z., Zhang, Y., and Yang, Q. Life Activity Modeling of News Event. Advances in Knowledge and Data Discovery 2012.

[15]. Cataldi, M., Di Caro, L., and Schifanella, C. Emerging topic detection on Twitter based on temporal and social terms evaluation. In Proceedings of the Tenth International Workshop on Multimedia Data Mining 2010.

[16]. S. Brin and L. Page, "Reprint of: The anatomy of a large-scale hypertex-tual web search engine,"Comput. Newts, vol. 56, no. 18, pp. 3825–3833, 2012.

[17]. E. Kwan, P.-L. Hsu, J.-H.Liang, and Y.-S. Chen, "Event identification for social streams is using keyword-based evolving graph sequences," in Proc. IEEE/ACM Int. Conf. Adv. Soc. Netw. Anal. Min., Niagara Falls, ON, Canada, 2013, pp. 450–457.

[18]. K. Kireyev, "Semantic-based estimation of term in formativeness," in Proc. Human Language Technol. Annu. Conf. North Amer. Chapter Assoc. Comput.Linguist., 2009, pp. 530–538.

[19]. J. Kleinberg, "Bursty and hierarchical structure in streams," Data Mining and Knowledge Discovery, vol. 7, no. 4, pp. 373–397, 2003.

[20]. M. Cataldi, L. Di Caro, and C. Schifanella, "Emerging topic detection on twitter based on temporal and social terms evaluation," in Proceedings of the Tenth International Workshop on Multimedia Data Mining. ACM, 2010, p. 4.
[21]. A. Dong, R. Zhang, P. Kolari, J. Bai, F. Diaz, Y. Chang, Z. Zheng, and H. Zha, "Time is of the essence: improving recency ranking using twitter data," in Proceedings of the 19th international conference on World wide web. ACM, 2010, pp. 331–340.
[22]. J. Weng and B.-S. Lee, "Event Detection in Twitter." ICWSM, vol. 11, pp. 401–408, 2011.
[23]. C. Li, A. Sun, and A. Data, "Tweet: segment-based event detection from tweets," in Proceedings of the 21st ACM international conference on Information and knowledge management. ACM, 2012, pp. 155–164.
[24]. L. Bolelli, b. Ertekin, and C. L. Giles, "Topic and trend detection in text collections using latent dirichlet allocation," in Advances in Information Retrieval. Springer, 2009, pp. 776–780.
[25]. T.Sakaki,M.Okazaki,andY.Matsuo,"EarthquakeshakesTwitterusers: real-time event detection by social sensors," in Proceedings of the 19th international conference on World wide web. ACM, 2010, pp. 851–860.