Flash Crowd Prediction in Twitter

Nitesh Kumar Singh
M.E student, Dept. Of CSE,
Anna University
Tamil Nadu, Chennai -25
mailme.nitesh24@gmail.com

Om Kumar C.U
Research Scholar, Dept. Of CSE,
Anna University
Tamil Nadu, Chennai -25
cuomkumar@gmail.com

Dr. Rajeshwari Sridhar
Asst. Prof (Sr.), Dept. Of CSE,
Anna University
Tamil Nadu, Chennai -25
rajisridhar@gmail.com

Abstract— **Computers have made their way into all realms of human actions. The emergence of high-speed networks and the popularization of mobile devices resulted in the increase of the number of accesses to online content on the Internet. Powerful servers have become inevitable as they support huge data processing. Flash crowd occurs due to popular content which attracts large crowd that demands immediate and instantaneous service which the server(s) find challenging. Research solutions till date propose dynamic substitution of resources or server(s) which is not cost effective. Our objective is to predict popular content that addresses the bottleneck's caused to server(s) by flash crowd. We develop a generic model that tracks trending topics in Twitter by extracting features through sentiment analysis, sarcasm analysis, trend analysis, emotional divergence, hash tag processing and predicts flash crowd through Naïve bayes classifier We benchmark the precision and recall of Naïve bayes classifier with KNN classifier to gauge its performance and throughput. The model depicts infectious rate which in turn helps in Flash crowd prediction in real time.**

Keywords— Feature extraction, sentiment analysis, sarcasm analysis, trend analysis, emotional divergence, hash tag processing, flash crowd.

I. INTRODUCTION

In this Era of Tera, everything needs to be digitized and automated. Computers have made their way into all realms of human actions in support of the task of digitizing and automating. Powerful servers have become inevitable as they support huge data processing [1-6]. Growth in hardware industries have led to production of high scale servers [10][11] that are equipped with adequate resources and bandwidth to respond to customer requests. This powerful servers has provoked the increase of the number of accesses to online content on the Internet. Due to the popularity of Open social networks [31][33], it is usual that a given content (e.g. a photo, a post or a video) has its popularity increased in a fast rate, becoming viral [37][38]. During times of newsworthy events, this high scale server are flooded with requests which cannot be handled by server's allocated resources and bandwidth thereby resulting in huge traffic[4] resulting in congestion and degradation in Quality of Service [12]. This is called "Flash Crowd Effect" [13][14][15]. The further implications are poor performance of websites, unnecessary utilization of client's bandwidth and computational capacity, transcontinental hops searching for resource, time delay in resource provisioning [7][8][9]. Our objective is to propose an efficient model that performs predictive analytics to analyze the popularity of online contents through Social media [35][36] and predict Flash crowd through patterns that deny unfruitful services in an efficient, scalable, resource optimized and cost effective manner.

We develop a generic model that tracks trending topics in Twitter [32][34] and does content/feature based popularity analysis through feature extraction, sentiment analysis, sarcasm analysis, emotional divergence and hash tag processing. Our model also analyze the network data in order to predict the popular content. To the best of our knowledge there is no such model that does content/features based analysis through such wide range of parameters and network data analysis for popularity prediction. The remaining sections of the paper is organized as follows; Section II describes various types of popularity prediction mechanism that can detect flash crowd. Section III describes the architecture of popularity prediction in twitter. Section IV discusses on training Naïve bayes classifier to predict viral and non-viral contents which in turn can be used to predict flash crowd. Section V Concludes the paper and proposes future work.

II.REVIEW OF LITERATURE

Sylvia [1] et.al and Dan Rubenstein [2] et.al in proposed Content Addressable Network which is basically a structured scalable peer to peer network overlay. Ion et.al proposed a distributed look up protocol to optimize search time of flash crowded documents in p2p overlay [3]. Jung et.al analyzed flash crowd caused by unauthorized people through DOS and DDOS attacks [4]. Xuan et.al proposed admission control mechanism to flash crowded applications [9]. Chenyu et.al addressed Flash crowd problems through intermediated caches that hold recently acquired popular contents [8]. Atajanov et.al developed anti flash crowd p2p proxies to avoid server being queued with huge number of requests [10]. Zhang et.al surveyed state of art solution to flash crowd through Content Deliver networks [12]. Broberg et.al developed and deployed MetaCDN a variant that combines various CDN to a single client [11]. Zaman et.al proposed his model that unveils information spreading in twitter [40]. The Wendell et.al studied flash crowd through CORAL an open CDN [13]. K.M. Prasad et.al has exclusively studied flash crowd through Botnets and has proposed counter measures through honeypots [14]. Naveed et.al published his research on controversial news spreading rate [33].

978-1-5090-4559-4/17/$31.00©2017IEEE

Bharti et.al has proposed ways of extracting sarcastic sentiment from tweets in real time [17]. Zhao et.al has developed seismic model to predict tweet popularity [27]. Jenders et.al discussed ways of analyzing and predicting viral tweets [28]. Palovics at.al proposed temporal prediction of retweet count [24]. Tsagkias et.al discussed the ways of predicting comments to online stories [29]. Zaman et.al discussed the ways of predicting information spreading in twitter [30]. Szabo et.al discussed in detail the ways of predicting popularity to an online content [31]. Lerman et.al used the content form social media to predict popular news [32]. Petrovic et.al proposed methods to forecast message propagation in twitter [34]. Asur [19] et.al and gupta [35] et.al discussed the ways of predicting trending events in twitter. Kupavskii et.al proposed the prediction of retweet cascade size of each tweet over a given interval of time [25]. Ma et.al proposed popularity through hash tags in twitter [37]. Ma et.al discussed ways of predicting viral content in a community structure of social media. Zaman et.al studied and proposed the Bayesian approach for popularity prediction [39]. Tatar et.al published a detailed survey on forecasting popularity of web content. Cheng et.al studied the ways of predicting cascading rate of a tweet [40].

III.INFERENCE

All the papers discussed above analyze flash crowd over a period of time and proposes a solution only after detecting it. The models used till date depict flash crowd either by running a thread on the server that alerts if the resource request reaches a threshold or through the exponential growth of request on a hot resource. The problem of popularity prediction in social networks has been widely studied through popularity, Retweet rate, viral contents and Flash crowd prediction through twitter is one of its kind.

We develop a popularity prediction mechanism that records the trending topics in twitter to efficiently predict flash crowd much ahead. Most of the related work on popularity prediction in twitter was carried out on a limited dataset with a limited number of settings. To our knowledge there are no studies which examine the contribution of individual features to popularity prediction. In this work we are studying the contribution of each individual feature for predicting the popularity of tweets.

Our proposed approach to popularity prediction as in Fig. 1 is based on a feature-based classification model in which we extract a set of features from tweets and classify them as popular/unpopular classes. To the best of our knowledge there is no prediction model that predicts the possibility of a flash crowd

IV. DESIGN

Understanding how users tweet and their motivations for tweeting is potentially important for predicting whether a tweet will be popular or not. In fact discovering what contents users choose to retweet can help to explain why a particular tweet becomes popular. Flash crowd can be depicted by identifying popular tweets. So by predicting popular contents from viral tweets [26] [27] [28] we can predict flash crowd. We propose a model that learns from a static twitter dataset of 10000 tweets and helps in predicting features of streaming tweets to be popular or not. Our definition of popular tweet revolves around a tweet's Retweet count. To allow a flexible definition of popularity we consider different Retweet-count, as a threshold for popularity call it popularity threshold [29] [30]. To evaluate the performance of our classifiers we compare the retweet probability with the different choices of evaluation such as Feature Extraction, Trend Score, Sentiment Analysis, Sarcasm, Emotional Divergence and Hashtag processing.

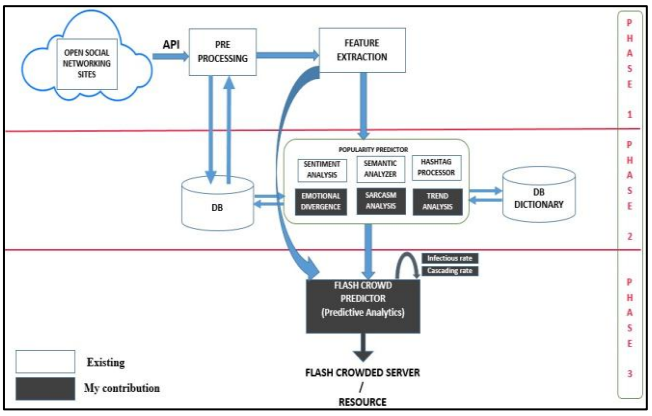

Fig. 1. Popularity Prediction

Sentiment Analysis:

Our Sentiment Analyzer [17] classifies tweets into three categories positive, negative and neutral. The goal of this analysis was to see whether retweet probability of tweets with positive sentiment undergo a different diffusion process than tweets with negative sentiment.

Trend Analysis:

Trend analysis [18] infers if the tweet is trending or not by checking the trending words in a tweet. This module tries to relate the inferred result to the retweet probability of a tweet.

Sarcasm

In the Sarcastic analysis we use Part of speech (POS) [16] tags that will analyze the situation sentiment of tweet and check whether it is contradicting the tweet sentiment. Our model also check for #sarcastic tags in tweet to analyze sarcasm. We analyzed how the inferred result relate to the retweet probability of a tweet. More specifically, for each tweet sarcastic and non-sarcastic label derived by sarcastic analyzer.

Emotional Divergence

Are emotionally diverse tweets more or less likely to be retweeted? This question extends the Sentiment one. It however shifts focus from the overall sentiment of a tweet to the emotional divergence encountered in the tweet. We have used the notion of emotional divergence as the

(normalized) absolute difference between the positive and the negative sentiment score delivered by Sentiment Analysis [27]. Whereas emotional polarity measures the overall emotion expressed in a text, emotional divergence measures the (extreme) span of expressed emotions. For example, according to Sentiment Analysis the sentence "I love hating you" will be classified as love=3 and hate=-4, hence resulting in binary emotional polarity b =−1 (negative). However, there might be high contrast in the emotional information of the used words (as is the case in the example) and emotional divergence is able to capture this effect.

Hashtag Processing

The Hashtag [19][20] performs a lookup search like(#happy, #sad, #Feelingexicted) in a tweet plays a prominent role in deciding if the tweet will be retweeted. The features listed above scores the tweet and classifies it to be either viral or non-viral. Viral tweets has more probability of becoming popular.

Tweet and Network Feature Extraction:

In addition to Tweet features like length of the tweet, number of mentions and URL's we also extract some additional features from the network of the user [25] who have posted the tweet. These features helps us to better exploit the information in the user's network which can potentially contribute to predict the popularity of tweet [22] [23][24]. This feature also helps us to predict the cascade [21] rate of tweet that will decide the popularity of tweet.

Table 1 Tweet & Network Features

Feature	**Type**	**Description**
Network Features		
Time	Discrete	Time of retweet
Number of Follower	Discrete	Number of follower's particular user have.
Tweet features		
Feature	**Type**	**Description**
Number of Retweet	Discrete	Number of times Retweeted
Length	Discrete	Total no. of words in a tweet
URL	Discrete	Number of URL's in a tweet
Mentions	Discrete	Number of mentions in a tweet

IV. EXPERIMENTAL ANALYSIS

We did our experiments of popularity prediction by training Naïve bayes classifier by extracting Static datasets of 10,000 tweets through twitter API. The trained classifier calculates scores of popularity thereby depicting cascading rate of tweet that helps in predicting Flash crowd.

Receiver Operating Characterestics (ROC) Curve:

To allow a flexible definition of popularity we consider different retweet-counts, as a threshold for popularity. The graph below tabulates the precision and recall value to different popularity thresholds. Receiver operating Characteristic curve in Fig. 3 of Naive Bayes Classifier presents the trade-off between the true positive rate and false positive rate. Closer the graph to its left border the more accurate the classifier.

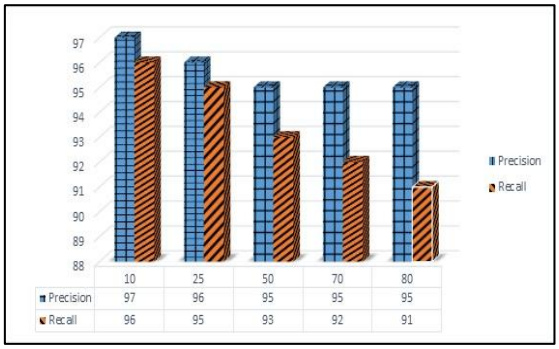

Fig. 2. Precision and Recall for different threshold value

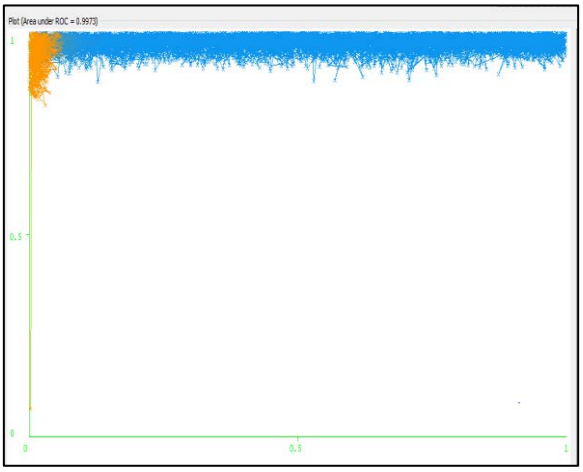

Fig. 3. ROC curve for Naive Bayes classifier

Comparision with KNN classifier:

We demonstrate the performance of our statically trained classifier through three new datasets in Fig. 4, 5, 6 (Two from the hot topics and 1 in general). Our classifier gauges the general dataset with better performance since it has a set of balanced (popular and unpopular) tweets when compared with other two popular datasets. Further to benchmark our popularity prediction we compare the results of our classifier with the baseline KNN calssifier. Precision and Recall greater than 85% in Fig. 7 indicates that the classifier is classifying very accurately. F-measure in Fig. 7 greater than 80% indicates the effectiveness of the corresponding class is based on classifier.

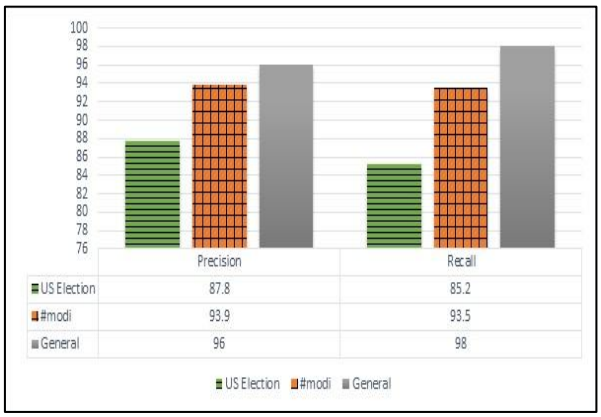

Fig. 4 Precision and Recall

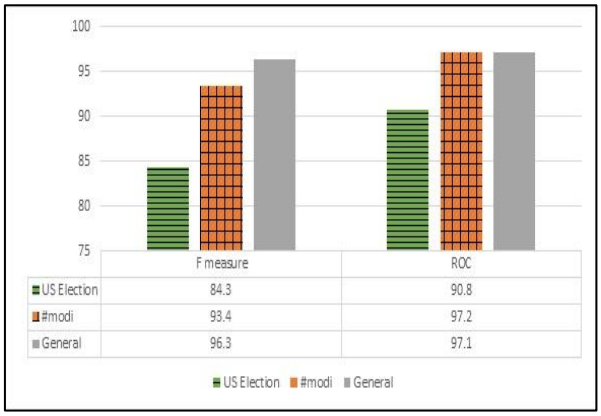

Fig. 5 F-Measure and ROC

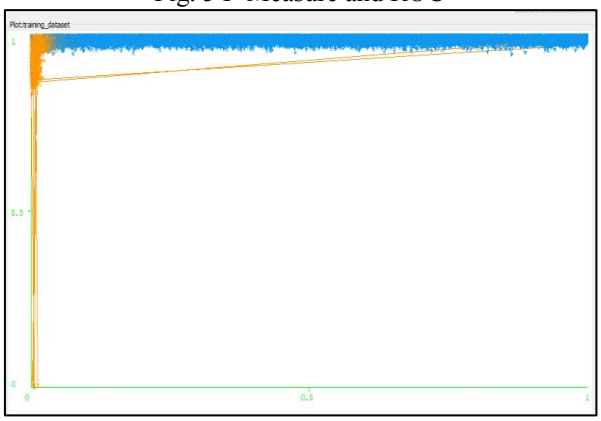

Fig. 6 ROC comparison b/n Naïve bayes and KNN classifier

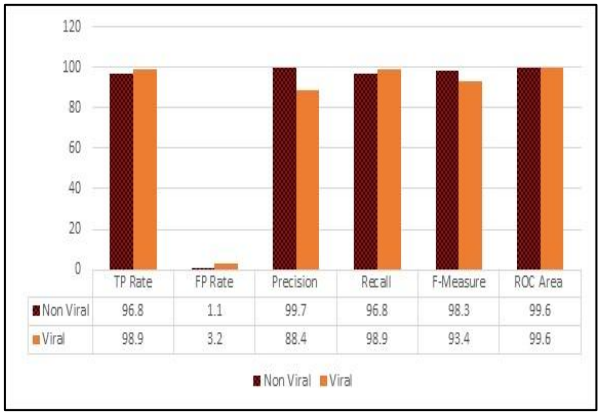

Fig. 7. Evaluation Metrics

Cascade Rate:

After predicting viral tweets, we move towards the network structure of the predicted viral tweet. We have used a seismic model[26] that assumes each post 'w' with a time dependent intrinsic cascade parameter p(w). In other words, p (w) models how likely the post 'w' is to be retweeted at time 't'. Fig. 8 shows the cascade rate at different time t. Fig. 9 will show the final output of flash crowd using cascading rate.

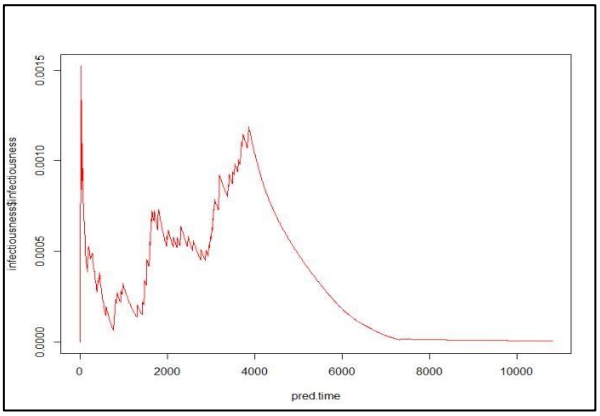

Fig. 8 Cascade rate of tweet

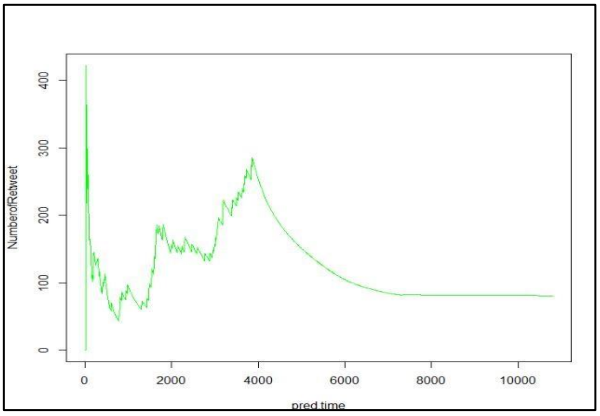

Fig. 9 Retweet ratio

We repeat our experiment with the interest to know if the selected features affects classifier's performance. It is clear from Fig. 10 that user and network features extracted gives a strong baseline. In addition, the combination of these features results in stable improvement of classifiers performance. We have projected Area under Cover (AUC) which actually summarizes the classifiers performance into a single quantity. Our results in Fig. 10 shows the consistency of classifier with its selected features.

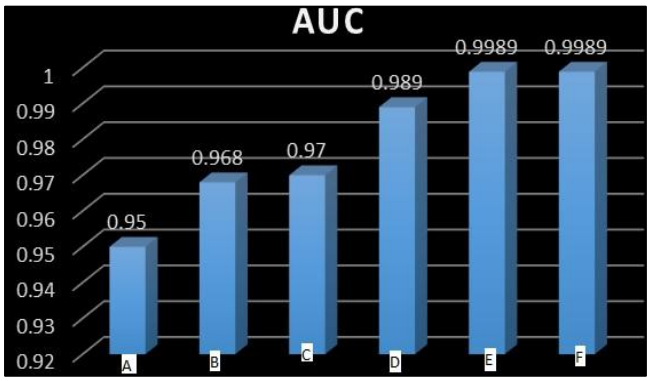

Fig. 10 Feature Combination for AUC

A-User & Network Features, B=A+ Trend, C= B+ Sentiment,

D= C+ Sarcastic, E= D+ Emotional Divergence, F=E+ Hashtag Processing

The below graph shows that after observing the cascade for 30 minutes the Absolute Percentage Error (APE) is 25%. And after observing the cascade for 60 minutes the APE is 22%. Observing the cascade for longer time decreases APE thereby increasing the efficiency.

This is because after observing a tweet for longer time we will have network data that will be very helpful in predicting cascade rate. But still considering network features for a long time reduces the chances of flash crowd persistence.

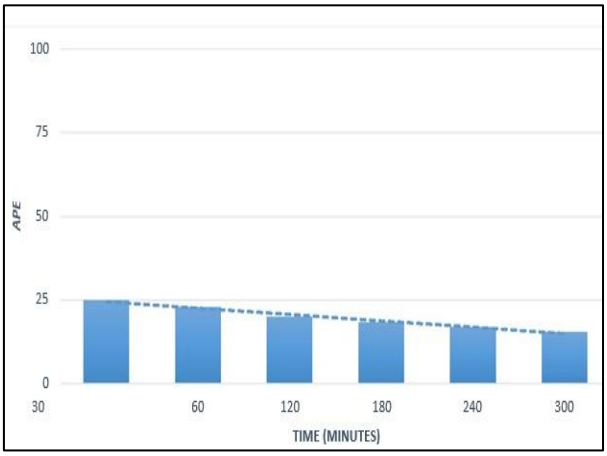

Fig. 11 Absolute Percentage Error for Flash Crowd Prediction

V.CONCLUSION AND FUTURE WORK:

In this work we proposed a popularity prediction mechanism that predicts popularity through statistical training naïve bayes classifier that classifies tweets by feature extraction, Hash tag processing, sarcastic analysis, sentiment analysis, Trend Analysis and Emotional Divergence and tries to predict whether the tweet be a popular tweet or not. We experimentally tested our approach by extracting 10000 tweets into 4 datasets using twitter streaming API. The statically trained classifier is compared with KNN classifier which exhibits efficiency of F- measure 80% precision and recall greater than 85%. The Absolute Percentage of error in Fig. 11 decreases over time. When time increases actually the chances of Flash crowd may decrease. So our future work is to extend the prediction in real time through streaming tweets which gives better chances of Flash crowd prediction.

REFERENCES

[1]. Sylvia Ratnasamy and Paul Francis, Mark Handley, Richard Karp, Scott Shenker, "A Scalable Content-Addressable Network", *ACM.SIGCOMM*, pp.161-172, 2001.

[2]. Dan Rubenstein and Sambit Sahu, "An Analysis of a Simple P2P Protocol for Flash Crowd Document Retrieval", pp.1-20, 2001.

[3]. Ion Stoica, Robert Morris, David Karger, M. Frans Kaashoek, Hari Balakrishnan, "Chord: A Scalable Peer-to-peer Lookup Service for Internet Applications", in IEEE/ ACM Transaction on Networking, pp.17-32, 2002.

[4]. Jaeyeon Jung, Balachander Krishnamurthy, Michael Rabinovich, "Flash Crowds and Denial of Service Attacks: Characterization and Implications for CDNs and Web Sites", in ACM, Proceedings of the 11th international conference of ACM World Wide Web, pp.293-304, 2002.

[5]. Tyron Stading, Petros Maniatis and Mary Baker," Peer-to-Peer Caching Schemes to Address Flash Crowds", in First Int. Workshop on Peer to Peer system, pp.1-11, MAR-2002

[6]. Ismail Ari, Bo Hong, Ethan L. Miller, Scott A. Brandt, Darrell D. E. Long, "Managing Flash Crowds on the Internet", *International Symposium of MASCOTS*, pp.246-249, 2003.

[7]. Angelos Stavrou, Sambit Sahu, "A Lightweight, Robust P2P System to Handle Flash Crowds", in IEEE Journal on selected areas of communication, pp.6-17, Jun.2003.

[8]. Chenyu, P. A. N., et al. "FCAN: Flash crowds alleviation network using adaptive P2P overlay of cache proxies", *IEICE transactions on communications*, pp.1119-1126, 2006.

[9]. Xuan chen and John heidemann, "Flash Crowd Mitigation via Adaptive Admission Control Based on Application-Level Observations", in ACM Transactions on Internet Technology, Vol. 5, No. 3, pp. 532–569, August 2005.

[10]. Ataianov, Merdan, Toshihiko Shimokawa, and Norihiko Yoshida."Autonomic multi-server distribution in flash crowds alleviation Network", *International Conference on Embedded and Ubiquitous Computing*. Springer Berlin Heidelberg, 2007.

[11]. Broberg, James, Rajkumar Buyya, and Zahir Tari. "MetaCDN: Harnessing Storage Clouds for high performance content delivery, " *Journal of Network and Computer Applications*, pp.1012-1022, 2009.

[12]. Zhang, Qi, Lu Cheng, and Raouf Boutaba. "Cloud computing: state-of-the-art and research challenges." *Journal of internet services and applications*, pp.7-18, 2010.

[13]. Wendell, Patrick, and Michael J. Freedman. "Going viral: flash crowds in an open CDN." *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*. ACM, pp.1-7, 2011.

[14]. Prasad, K. Munivara, M. Ganesh Karthik, and ES Phalguna Krishna. "An Efficient Flash Crowd Attack Detection to Internet Threat Monitors (ITM) Using Honeypots." *Advances in Computing and Information Technology*. Springer Berlin Heidelberg, pp.595-610, 2013.

[15]. de Paula, Ubiratam, et al. "Detecting and Handling Flash-Crowd Events on Cloud Environments", in press Transaction on Web, *arXiv preprint arXiv:1510.03913* (2015).

[16]. Bharti, S. K., et al. "Sarcastic sentiment detection in tweets streamed in real time: a big data approach." Digital Communications and Networks, pp.108-121, 2016.

[17]. Saif, Hassan, et al. "Semantic patterns for sentiment analysis of Twitter." International Semantic Web Conference. Springer International Publishing, 2014.

[18]. Asur, Sitaram, et al. "Trends in social media: Persistence and decay." *Available at SSRN 1755748* (2011).

[19]. Kong, Shoubin, et al. "Predicting bursts and popularity of hashtags in real-time." Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval. ACM, 2014.

[20]. Cui, Peng, et al. "Cascading outbreak prediction in networks: a data-driven approach." *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2013.

[21]. Cheng, Justin, et al. "Can cascades be predicted?." Proceedings of the 23rd international conference on World wide web. ACM, 2014.

[22]. Gao, Shuai, Jun Ma, and Zhumin Chen. "Modeling and predicting retweeting dynamics on microblogging platforms." *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. ACM, 2015.

[23]. Pálovics, Róbert, Bálint Daróczy, and András A. Benczúr. "Temporal prediction of retweet count." Cognitive Infocommunications (CogInfoCom)IEEE, 2013.

[24]. Kupavskii, Andrey, et al. "Prediction of retweet cascade size over time." Proceedings of the 21st ACM international conference on Information and knowledge management. ACM, 2012.

[25]. Vasconcelos, Marisa, Jussara M. Almeida, and Marcos André Goncalves. "Predicting the popularity of micro-reviews: A Foursquare case study." *Information Sciences*, pp.355-374, 2015.

[26]. Zhao, Qingyuan, et al. "Seismic: A self-exciting point process model for predicting tweet popularity." Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2015.

[27]. Jenders, Maximilian, Gjergji Kasneci, and Felix Naumann. "Analyzing and predicting viral tweets." Proceedings of the 22nd International Conference on World Wide Web. ACM, 2013.

[28]. Tsagkias, Manos, Wouter Weerkamp, and Maarten De Rijke. "Predicting the volume of comments on online news stories." *Proceedings of the 18th ACM conference on Information and knowledge management*. ACM, 2009.

[29]. Zaman, Tauhid R., et al. "Predicting information spreading in twitter." *Workshop on computational social science and the wisdom of crowds, nips*. Vol. 104. No. 45. Citeseer, 2010.

[30]. Szabo, Gabor, and Bernardo A. Huberman. "Predicting the popularity of online content." *Communications of the ACM* 53.8 (2010): 80-88.

[31]. Lerman, Kristina, and Tad Hogg. "Using a model of social dynamics to predict popularity of news." *Proceedings of the 19th international conference on World wide web*. ACM, 2010.

[32]. Naveed, Nasir, et al. "Bad news travel fast: A content-based analysis of interestingness on twitter." *Proceedings of the 3rd International Web Science Conference*. ACM, 2011.

[33]. Petrovic, Sasa, Miles Osborne, and Victor Lavrenko. "RT to Win! Predicting Message Propagation in Twitter." *ICWSM*. 2011.

[34]. Gupta, Manish, et al. "Predicting future popularity trend of events in microblogging platforms." *Proceedings of the American Society for Information Science and Technology* ,pp.1-10,2016.

[35]. Bandari, Roja, Sitaram Asur, and Bernardo A. Huberman. "The pulse of news in social media: Forecasting popularity", 2012.

[36]. Ahmed, Mohamed, et al. "A peek into the future: predicting the evolution of popularity in user generated content." *Proceedings of the sixth ACM international conference on Web search and data mining*. ACM, 2013.

[37]. Ma, Zongyang, Aixin Sun, and Gao Cong. "On predicting the popularity of newly emerging hashtags in twitter." *Journal of the American Society for Information Science and Technology* ,pp.1399-1410, 2013.

[38]. Weng, Lilian, Filippo Menczer, and Yong-Yeol Ahn. "Virality prediction and community structure in social networks", 2013.

[39]. Zaman, Tauhid, Emily B. Fox, and Eric T. Bradlow. "A Bayesian approach for predicting the popularity of tweets.", pp. 1583-1611,2014.

[40]. Tatar, Alexandru, et al. "A survey on predicting the popularity of web content." *Journal of Internet Services and Applications,* 2014.