Tweet Specific Extractive Summarization Framework towards Trending Topic Analysis

Halima Banu S
Department of Computer Science and Engineering
College Of Engineering Guindy, Anna University
Chennai, India
halimabanu91@gmail.com

Chitrakala S
Department of Computer Science and Engineering
College Of Engineering Guindy, Anna University
Chennai, India
au.chitras@gmail.com

***Abstract*— Twitter establishes itself as a critical element to any social networking based applications serving itself as a rich information delivering platform. However, some users, especially new users, often find it difficult to understand trending topics in Twitter when confronted with overwhelming and unorganized tweets. Previously, there has been attempts to provide a short snippet to summarize a topic but, this does not scale up to user's expectation as it does not provide any *analyzed summary of tweets*. This work aims to develop a Tweet Specific Extractive Summarization system performing Trending Topic Analysis. This system analyzes trending topics through Topic based sentiment classification which summarizes the public views on selected trending topics and generates extractive sub summaries of topics over the time period using novel Tweet Feature Graph Model (TFGM). Tweet Specific Extraction Summarization framework differs from the traditional summarization in few aspects. First, conflicting summary generation could be avoided with sentiment classification enhanced by common and tweet specific feature extraction thereby sorting the data into separate sentiment corpus. Second, volume-based followed by topic modelled approach of detecting sub topic in the corpus help detect subtopics under the trending topic more efficiently. Finally, Summary generation is accomplished using the Tweet Feature Graph Model (TFGM) which incorporates tweet specific salient features. This model increases relevancy of tweets in content selection phase which in turn contributes to the increase in quality of the summaries generated.**

***Index Terms*— Extractive Summarization, Trending Topic Analysis, Sub-Topic Detection, Tweet Feature Graph Model (TGFM)**

I. INTRODUCTION

In this era it has become evident that Twitter has become a mainstream medium for dissemination of messages and the public discussion of news and events. The unprecedented rate of tweets sets a big obstacle for efficient information acquisition. It is impossible for a user to get an overview of important topics on Twitter just by merely reading an exhaustively huge amount of tweets every day. In addition, because of information redundancy and the informal writing style, it is time consuming to find useful information about a topic from a huge volume of tweets. This tremendous number of tweets raises a needful framework called summarization as the key to facilitating the requirements of topic exploration, topic diffusion, and search from hundreds of thousands of tweets. Specifically, a summary that provides representative information of topics with no redundancy and overlapped content is desired.

To obtain a complete analysis of the content i.e. tweets it is required to perform sentiment classification but summarization applications lack this component and as a result produce conflicting summaries. Topics discussed in Twitter are very diverse and unpredictable. Sentiment classifiers always dedicate themselves to a specific domain or topic. Namely, a classifier trained on sentiment data from one topic often performs poorly on test data from another. Most of the recent applications for sentiment analysis deal with a single topic under interest. Another concern regarding sentiment classification is if carried out as a supervised approach it will incur heavy manual effort in annotation. Major focus of this work lies in: (1) Handling avoidance of generation of conflicting summaries (2) Performing Topic Evolution in summaries generated (3) Building a Tweet specific summarization framework.

Rest of the paper is organized as follows: In Section 2 we discuss the literature survey pertaining to Sentiment Classification and Twitter based Extractive Summarization. System architecture and details about the proposed model in described in Section 3. Experimental Results are discussed in Section 4 followed by Conclusion and Future works in Section 5.

II. LITERATURE SURVEY

A. Twitter Based Extractive Summarization

The popularity of micro-blogging services, such as Twitter has caught increasing attention from worldwide researchers. There exist some pioneering researches working on Twitter summarization.

G.Mane et all [1] proposed a combined approach of Phrase Reinforcement algorithm along with Word Sense

978-1-5090-2552-7

Disambiguation and Textual Entailment techniques for generating one line summary. Phrase Reinforcement algorithm aimed at constructing a graph which helps in identifying the most commonly occurring phrases for a central topic by simply searching for the most weighted set of paths through the graph. This methodology lacked temporal nature of summaries and created coherence issues in the summary generated. Moreover
since it is intended to generate one line summaries it does not capture the specific feature of a tweet and this would lead to ill -formed summaries based on tweets.

D.Wen et all [2] accomplished Summarization using a non-parametric Bayesian model applied to Hidden Markov Models. A novel observation model was designed to allow ranking based on selected predictive characteristics of individual tweets. Major focus was to investigate the possibility of using a temporal probabilistic data model known as Hierarchical Dirichlet Process Hidden Markov Model (HDP-HMM) to process a stream of tweets pertaining to a single subject and cluster the tweets into groups or rankings based on the value of the individual tweets. But Summaries generated did not project the temporal nature.

F.Liu et all [3] proposed a Concept Based optimization framework for topic Summarization using Integer Linear Programming. Target data comprised of original and normalized tweets along with web content relevant to the topic. The focus was not on developing new summarization systems but rather utilizing and integrating diverse text sources to generate more informative summaries but there were lack of series of sub events identification to show topic evolving process.

B.O'Conor et all [4] explained Twitter topics by presenting a simple list of messages. An exploratory search application for Twitter called TweetMotif grouped messages by frequent significant terms and retrieved results were further narrowed down through a search interface. But the system lacked temporal aspect in summaries generated and topic development was not observed since sub topic detection process was not employed.

D.Gao et all [5] generated sequential summaries on a topic using Stream and Semantic based approaches for sub topic detection. But the approach fell short of readability of summaries generated. System failed to capture the opinion expressed in the data thus leading to conflicting summaries. Also, sub topic detection mechanism needs more enhancement.

B. Sentiment Classification

One of the major issues of Sentiment classification using machine learning approaches is the need for labeled training data. Twitter dataset lack labeled data which often poses a huge effort in manual annotation.

H.Rui et all [6] used supervised algorithms namely Naïve Bayes Classifier and Support Vector Machine Classifier for sentiment classification to investigate if Twitter WOM affects movie sales and if so the approach involved in it. Contributions included measuring the impacts of WOM from people with different degrees of social connectivity exhibited in a social network. But the system incurred heavy labor work in labeling the huge training data set.

Furthermore twitter data are very diverse; therefore to train a universal classifier was near impossible. Some works have been done for cross domain sentiment analysis for review datasets like in [7], S.J.Pan et all used Spectral Feature Alignment algorithm (SFA) for classifying sentiment polarity, or bridging the gap between the domains and to align domain-specific words from different domains into unified clusters. System significantly outperformed previous approaches to cross-domain sentiment classification but could not be applied for twitter dataset.

Xiaojun Wan [8] used Co training algorithm for training and SVM Classifier was used for sentiment analysis. Major focus was on cross lingual sentiment classification which leveraged English corpus for Chinese sentiment classification. But the approach could not suite well for the nature of tweets.

S.Lui et all [9] performed semi supervised multi class model for topic adaptive sentiment classification. The classifier was initially built using common features which did not adapt well for cross domains. Semi supervised learning was used to adaptively learn the build a topic adaptive classifier. But the system needed topic labeled data in order to apply the respective classifier model. Obtaining labelled data poses a major constraint which in turn reduces the adaptability of this system to any kind of data to be analyzed.

III. SYSTEM ARCHITECTURE

Trending Topic Analyzer based on tweet specific summarization framework is implemented through topic adaptive sentiment classification and multi tweet extractive summarization. This work aims at generating chronologically ordered sub summaries which throws light in understanding the topic evolution of the trending topics. Topic under study is assumed to contained several hidden sub topics which are revealed using pipelined sub topic detection models. Summary generation is accomplished using a proposed graph based model named Tweet Feature Graph Model (TFGM). This framework delivers a complete analyzed and relevant tweet specific summaries.

A. Target Data Collection and Tweet Pre-processing

Target data collection is performed by extracting the trending topic on a region basis. Only those topics which involves a topic development of user focus change will be considered for summarization in this work. Based on the selected topics tweets are extracted from the Twitter and stored in database. Necessary preprocessing such as URL removal, Slang Word Replacement, Non English Tweet Filter, Stemmer and Stop Word Removal is performed to prepare the target useable data. Along with the preprocessing phase, the proposed system also handles other language tweet translation which is essential to prevent discarding of public opinion about the trending topic.

B. Topic Based Sentiment Classification

Our sentiment classification is inspired by the method proposed by Shenghua Liu et al. in [9] for labelling the corpus with their corresponding sentiments. Topic based sentiment

978-1-5090-2552-7

classification is divided into two phases namely Feature Extraction and Model Building.

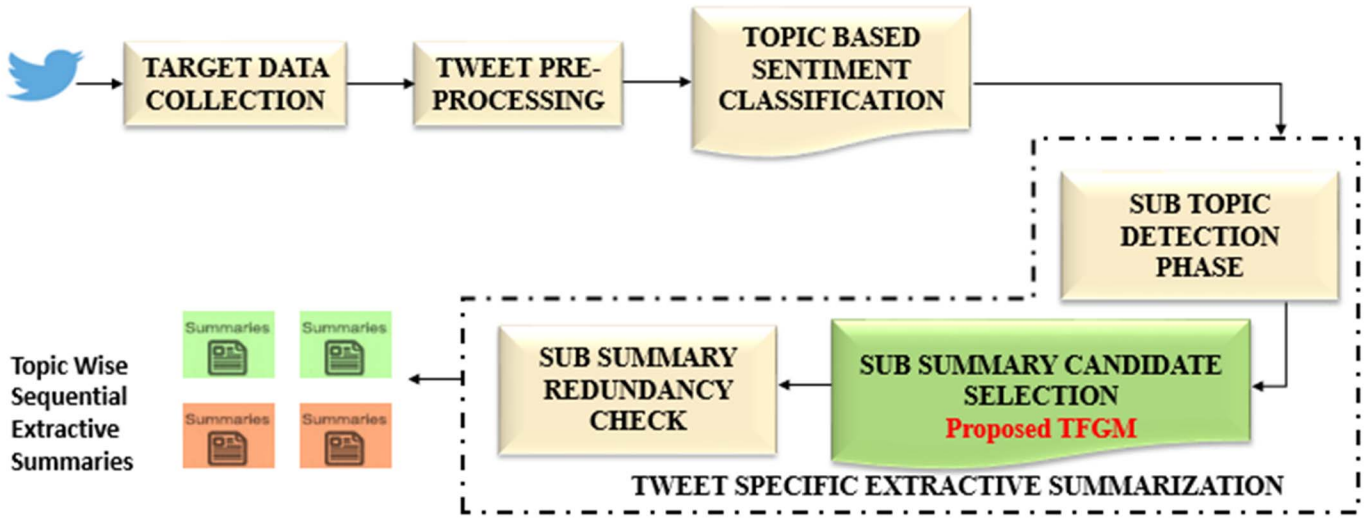

Fig.1 System Architecture.

Feature extraction involved in this system creates a significant impact in the performance of sentiment classification task. Tweets have a special nature unlike normal English sentences like @ symbol used to refer to a user, emoticons expressed in the tweets etc. Incorporating such features while performing feature extraction enhances the overall process. Features extracted could be classified into two broad categories namely Common Features and Tweet Specific Features. With the extracted features classifier model is built based on each topic without exhaustively labelling the training set. Support Vector Machine is used as the classifier to perform the semi supervised classification task. The classifier model is built using a collaborative training approach where the unlabelled data and the features extracted will be iteratively augmented with the initial labelled set. Figure 1 above shows the system architecture depicting the proposed summarization model.

C. Tweet Summarization Framework

Topic evolution in summarization is achieved using two-step pipelined phases namely *Sub Topic Detection* and *Sub Summary Generation* [5]. Tweets are prone to contain noisy data due to various reasons such as limited length of a tweet text, informal language used by the user tweeting etc. *Sub Summary Redundancy Check* phase handles redundant content along with fairness in content selection.

(i) Sub Topic Detection Phase

Sub Topic Detection process in the proposed framework involves a Volume based approach followed by a Topic Modelling approach called the Foreground-Dynamic Topic Modelling (F-DTM).

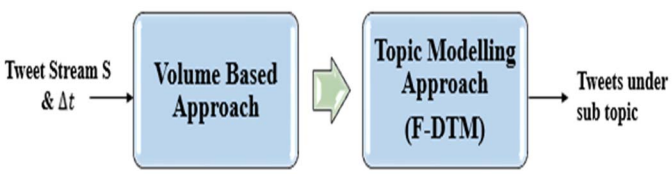

Fig.2 Sub Topic Detection Phase

Tweet stream that is collected may contain many irrelevant data which may tamper the quality of the extractive summary generated. This raises a need to clean and extract only the relevant corpus. Noisy and irrelevant tweets given a trending topic is handled by a bi-level process of executing Volume based approach and; the output from the previous step i.e. the tweet set collected will be fed into the novel topic modelling approach.

(ii) Sub Summary Generation Phase

Each tweet is limited to 140 characters which poses a challenge to extract most significant tweets from the corpus mixed with noisy data. Various graph based tweet selection proved to be more efficient than other form of approaches. In this work a novel model named Tweet Feature Graph Model (TFGM) is proposed which captures the salient nature of tweets leading to more relevant content selection.

Algorithm 1: Tweet Feature Graph Model **(TFGM)**

Input *: Tweets under sub topics -* t_s
Output *: Score tweets for each sub topic,* ret

Initialize the set of tweet vectors t_s

for each tweet $t \in t_s$ *do*
  get a feature vector $\vec{t} = repr(t)$
  add $\vec{t}$ *to* ts_v: $ts_v \cup \{\vec{t}\}$
end for

begin
  Compute modified weights using retweet count and followers count as per eq. (1) and eq. (2)
  Construct $graph(ts_v)$
  Compute salience scores: $sc = score(gr)$
  Select tweets: $ret = select(sc, ts)$
  return ret
end

Function $repr(t)$
represent each tweet into bag-of-words vector

Function $\boldsymbol{graph(ts_v)}$

compute transition probability p(i.j) as in eq. (3)

Function $\boldsymbol{score(gr)}$

compute salience score for each tweet as per eq. (4)

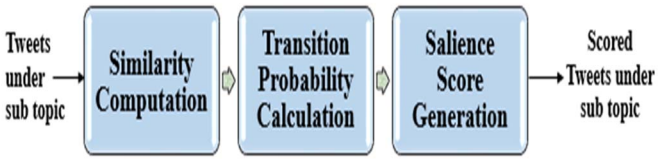

Fig.3 Sub Summary Generation Phase

Steps involved in Social feature graph model is depicted in Figure 3 above. Initially the tweets will be represented as a tweet vector. Cosine similarity between tweets in the corpus is calculated and the weights of each edge which incorporates special nature of tweets namely: *Retweet number and Followers count* are computed as in eq. (1) and eq. (2).

$$w(i,j) = \frac{sim(\vec{t_i},\vec{t_j})}{\sum_j sim(\vec{t_i},\vec{t_j})} \frac{a_j}{a_{j'}} \quad (1)$$

$$a(j) = retw(j).foll(j) \quad (2)$$

Transition probability between edges are calculated according to eq. (3).The transition probability *p(i,j)* in eq. (3) is sufficiently different from *p(j,i)* because of the different normalization factor in the denominator.

$$p(i,j) = \begin{cases} \frac{w(i,j)}{\sum_j w(i,j)} & if \ \sum_j w(i,j) \neq 0 \\ 0 & otherwise \end{cases} \quad (3)$$

Finally, the salience score for each tweet will be calculated as shown in Figure 3 according to eq. (4) where s_i is the salience score of the i^{th} vertex, λ is the damping factor and V is the number of tweets for summarization.

$$s_i = \lambda . \sum_{j \neq i} s_j . p(i,j) + (1-\lambda).1/|V| \quad (4)$$

The Figure 4 below displays the salience scores for sample number of tweets. This score will decide whether the respective tweet will be candidate for selection and the highly ranked tweets are observed to increase relevancy of content selection.

serial_no	score
30043	316.1641767250335
30044	316.1623268194373
30045	315.93790368257623
30046	315.936750426555
30047	316.0251532317782
30048	316.0233692994774
30049	416.1695638225113
30050	207.4529521340216
30051	207.44848385800628
30052	208.43451288701914
30053	208.43054945445382
30054	210.04039480224452
30055	210.0255918026414
30056	208.28548587952952
30057	207.45368714234897
30058	208.4288003451847

Fig.4 Sub Summary Candidate Selection

(iii) Sub Summary Redundancy Check

Whether a tweet is selected as a representative tweet depends on two factors: its *salience score* and its *similarity* to the already selected tweets. Specifically, a tweet is chosen if it is the candidate with the greatest salience score and its similarity to any selected tweet is below a threshold. No matter whether the most salient candidate is chosen or not, it will be removed from the candidate. This selection process repeats until M tweets are chosen or the candidate set is empty. Algorithm below details about the flow involved. ε in the algorithm is the threshold which has to be satisfied for the tweet to be selected.

Algorithm 2: Sub Summary Redundancy Check

Input ***: Tweets under sub topics -*** $\boldsymbol{t_s}$
Output ***: Score tweets for each sub topic,*** $\boldsymbol{ret}$

Initialize the set of selected tweets
$sel: sel = \emptyset$
Initialize set of candidates
$cand: cand = \{(\vec{t}, score)\}$
while $|sel| < M \ and \ |cand \neq 0|$ *do*
select the most salient
$t^*: t^* = \arg max_{t \epsilon cd} \, t.score$
remove t^* *from* $cand: = cand - \{t^*\}$
if $\forall s \in sel, sim(s.\vec{t}, t^*.\vec{t}) < \varepsilon$ *then*
select $t^*: sel = sel \cup \{t^*\}$
end if
end while
check userGroup diversity
return sel

To ensure that the analysed result is not skewed towards a smaller group of users, a check is performed adhering to user group diversity.

978-1-5090-2552-7

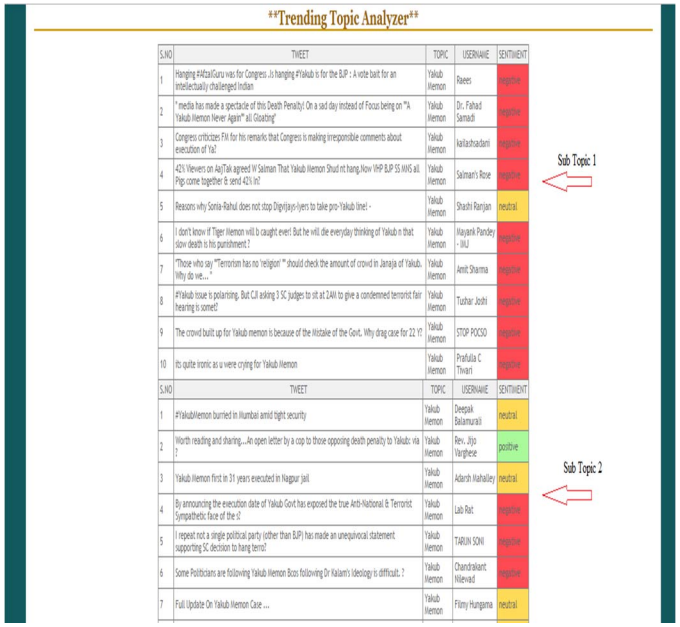

Fig.5 Sample Extractive Sub Summaries Generated via proposed Summarization Framework

VI. EXPERIMENTAL RESULTS

Twitter API allows to track and download the trending topics given the WOEID (Where On Earth Identifier) of a region. Based on the selection criteria, only those topics which involves topic development or user focus change will be considered and hence downloaded. Tweets from July 2015 to September 2015 are collected. Classifier's performance is measure using Weka[11] tool.

Performance of extractive sub summaries generated are measured using two popular metrics for summarization namely Coverage and Novelty. The coverage is defined using eq. (5):

$$\text{Coverage} = \frac{1}{|D^H|} \cdot \sum_{d_i \in D^H} \frac{1}{w_{ij}} \frac{\sum_{d_j \in D^S} \sum_{N-gram \in d_i^H, d_j^S} \sum Count_{match}(N-gram)}{\sum_{d_j \in D^S} \sum_{N-gram \in d_j^S} \sum Count\,(N-gram)} \quad (5)$$

$$where,\ w_{ij} = \begin{cases} |j-i|+1, j \neq i \\ 1 \qquad\quad , j = i \end{cases}$$

where, D^H and D^S are human and system generated sub summaries, i and j are the index of the sub summary. w_{ij} is the weight added corresponding to the matches between the system and human generated sub summaries. Novelty is defined using eq. (6):

$$Novelty = \frac{1}{|D|-1} \sum_{i>1} (I_{d_i} - I_{d_i, d_{i-1}}) \quad (6)$$

where, I_{d_i} is the Information Summary and $I_{d_i, d_{i-1}}$) is the overlapped information. This is used to calculate the average increment of information content of two adjacent sub summaries.

TABLE 1

COMPARISON WITH BASELINE SYSTEMS

Trending Topic	Metric	Heuristic-baseline	Human Generated	**Proposed TFGM based Summarization Framework**
#Yakub Memon	Coverage	2.15	4.20	3.55
	Novelty	2.26	4.95	4.10
#Black Day For Indian Democracy	Coverage	2.10	4.05	3.42
	Novelty	2.13	4.90	4.00

Performance of the proposed framework is compared against baseline systems and human generated summaries and the results are depicted in Figure 6.

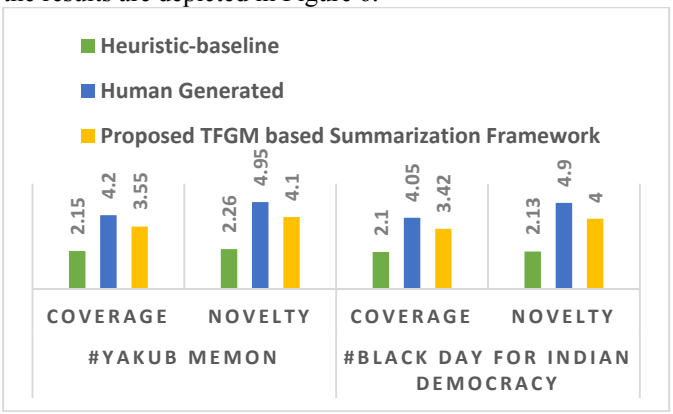

Fig.6 Comparitive Analysis

V. CONCLUSION

Thus the proposed Tweet Specific Extractive Summarization Framework for achieving a Trending Topic Analysis has been implemented and is able to analyse twitter trending topics in a constructive and accurate manner. The system extracts the most significant and relevant tweets to generate extractive sub summaries using a novel graph based model named Tweet Feature Graph Model (TFGM) incorporating the salient features of the tweets. Extractive Sub Summaries generated are free from redundancy and tagged with its respective sentiment label. The system can be further enhanced by normalizing the selected significant tweets for better understanding and can be modelled to generate abstractive summaries of the topic which also contributes to the readability of the summary.

REFERENCES

[1] G.Mane, A.Kulkarni,"Twitter Event Summarization Using Phrase Reinforcement Algorithm and NLP Features", *International Journal of Advanced Research in Computer and Communication Engineering* Vol. 4, Issue 5, May 2015,pp. 427-430

[2] Wen, Dunwei, and Geoffrey Marshall. "Automatic Twitter Topic Summarization." *Computational Science and Engineering (CSE), 2014 IEEE 17th International Conference on*. IEEE, 2014, pp-207-212

[3] Liu, Fei, Yang Liu, and Fuliang Weng. "Why is SXSW trending?: exploring multiple text sources for Twitter topic summarization." *Proceedings of the Workshop on Languages in Social Media*. Association for Computational Linguistics, 2011,pp.66-75

[4] O'Connor, Brendan, Michel Krieger, and David Ahn. "TweetMotif: Exploratory Search and Topic Summarization for Twitter." *ICWSM*. 2010,pp.384-385

[5] D.Gao, W.Li, X.Cai, R.Zhang, and Y.Ouyang, "Sequential Summarization: A Full View of Twitter Trending Topics", *IEEE*

Transactions on Audio, Speech, and Language Processing, Vol. 22, no. 2,pp. 293-302,Feb. 2014

[6] Rui, Huaxia, Yizao Liu, and Andrew Whinston. "Whose and what chatter matters? The effect of tweets on movie sales." *Decision Support Systems*55.4 (2013): pp.863-870.

[7] Pan, S. J., Ni, X., Sun, J. T., Yang, Q., & Chen, Z. (2010, April). "Cross-domain sentiment classification via spectral feature alignment." In*Proceedings of the 19th international conference on World wide web* (pp. 751-760). ACM

[8] Wan, Xiaojun. "Co-training for cross-lingual sentiment classification." *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*. Association for Computational Linguistics, 2009.pp.235-243

[9] S.Liu, X.Cheng, F.Li, and F.Li,"TASC: Topic-Adaptive Sentiment Classification on Dynamic Tweets", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 27 ,no. 6,pp. 1696 - 1709,Jun.2011

[10] S.Tan, Y.Li, H.Sun, Z.Guan, and X.Yan, "Interpreting the Public Sentiment Variations on Twitter", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 26, no. 5, pp. 1158-1170, May. 2014

[11] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann , and I. H. Witten, "The weka data mining software: An update," ACM SIGKDD

978-1-5090-2552-7