IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-3, Issue- 3 E-ISSN: 2347-2693

REAL TIME EXTRACTION AND PROCESSING OF SOCIAL TWEETS

B. M. Bandgar[1*] and Binod Kumar[2]

Dept. of Computer Science, Karpagam University, Coimbatore-641021, India
[2] *JSPM's Jayawant Technical Campus*, Tathwade, Pune-411033, India
bapuraob@yahoo.com, binod.istar.1970@gmail.com

www.ijcseonline.org

Received: Feb/09/2015 Revised: Feb/22/2015 Accepted: Mar/10/2015 Published: Mar/10/2015

Abstract— Twitter has become one of the most popular micro-blogging platforms recently. Millions of users can share their thoughts and opinions about different aspects and events on the micro-blogging platform. Therefore, Twitter is considered as a rich source of information for decision making and sentiment analysis. Obtaining the real time tweets on the particular topic is one of the challenging tasks. There are numbers of related publications, but which have limitations; also their methods are not much clear and mostly based on Linux based system and uses integrated tools, which is most complex job. Therefore, in this research paper we develop the indigenous Windows based user friendly application in Java to extract, process and classify the real time social network tweet. The tweets are processed for removal of hash, tags and URL and removed the stop words from sentence and tried to detect, analyze the abbreviations or slangs. The meaningful real time tweets are obtained and used for sentimental analysis.

Keywords— *Extraction of realtime tweets, processing tweets*

I. INTRODUCTION

The emergence of social media has given web users a venue for expressing and sharing their thoughts and opinions on different topics and events. Twitter, with nearly 600 million users and over 250 million messages per day, has quickly become a gold mine for organizations to monitor their reputation and brands by extracting and analyzing the sentiment of the Tweets posted by the public about them, their markets, and competitors. Sentiment analysis over Twitter data and other similar micro-blogs faces several new challenges due to the typical short length and irregular structure of such content. The following are some of the challenges faced in sentiment analysis of Twitter feeds

- Named Entity Recognition (NER) — NER is the method of extracting entities such as people, organization and locations from twitter corpus.
- Parsing - The process of identifying the subject and object of the sentence. The verb and adjective are referring to what?
- Sarcasm - What does a verb actually stand for? Does 'bad' mean bad or good?
- Sparsity-Insufficient data or very few useful labels in the training set.
- Twitter abbreviations, poor spellings, poor punctuation, poor grammar, incomplete sentences.
- The accuracy of tweets classification as compared to human judgments.

Bifet, A. and Frank, E. [2] proposed a data mining technique used for sentiment knowledge in twitter data streams. The proposed algorithm focuses on classification of data streams and performs sentiment analysis in real time. The evaluation of results is verified using a sliding window kappa statistics that works for constantly changing data streams. Only a small number of tweets (177 negative and 182 positive) were used to test the accuracy. This is a very small number of tweets to make any judgment about the proposed technique. Only tweets containing an emoticon were considered; which is also a very small portion of overall tweets. The paper uses a balanced dataset which is not a sample of real-time Twitter stream which is normally unbalanced. Bifet, A., Holmes, G., and Pfahringer, B. [3] discussed the handling of tweets in real-time.

The research paper introduced a system, MOATweetReader, which processes the tweets in real-time despite of their dynamic nature. The system performs two functions: First, it detects the changes in term frequencies and second, it performs sentiment analysis in real-time. Some applications of the proposed framework have also been discussed in frequent item mining and sentiment analysis. The paper shows a correlation between the twitter sentiments and the Toyota crisis and successfully claims that the MOA-Tweet Reader tool could have identified the crisis coming. Argamon, S. et al.[4] used a supervised learning algorithm for determining complex sentiment-related attributes. These attributes are classified as attitude type and force. The results when averaged showed that the Naïve Bayes algorithm is the best among the lot. SVM also dominates the results. The

© 2015, IJCSE All Rights Reserved

proposed algorithm is well suited where lexicons need to be generated from the scratch. Fu, X. et al.[5] presented a method for semantic extraction using information theoretic co-clustering. The proposed algorithm is based on implicit associations within evaluated features, within evaluated semantic words and between evaluated features and semantic words. A feature semantic word matrix represents the co-occurrence relationships of feature words and semantic words. Then, co-clustering algorithm is applied on this matrix for clustering of these evaluated features. The Chinese language dataset are used for testing. No detailed information is available about how the manual analysis of the online review was conducted. The experimental results showed 78% accuracy.

Here we extracted the real time tweets related to the social network, on which real time feed backs can be obtained on the particular event happening the particular geographical area from the different sources such as twitter news websites etc. Further we tried to process the raw tweets sothat the same processed tweets can be used for the sentimental analysis.

II. EXPERIMENTAL DETAILS

The twitter4j API version 1.01 downloaded from [7] and created application on the web site[8]. The token and access keys obtained, which is required for the extraction of the tweets. An indigenous application is developed using NetBeans IDE 8.0 as platform in Java based on windows operating system and extracted the real time tweets from the twitter website and news websites. The extraction is repeated for multiple times to obtain more number of tweets. The extracted raw tweets pre-processed for removal of URL and removal of all the private usernames identified by @user and the hash tags identified by the # symbol using a regular expression in Java. Lastly we removed all the special characters excluding the emoticons. The pre-processed tweets used to detect, analyzes abbreviations or slang for obtaining meaning full message.

III. RESULTS AND DISCUSSIONS

The real time tweets were obtained on the Monday, 16th June 2014 at 20:52:200 IST 2014 for the keyword 'Narendra Modi' using indigenously developed application. The sample of resulted raw tweets is given in the table 1. The cleaning of the tweets has been done by different methods. Initially pre-processing of the raw tweets is carried out, only useful sentences are sorted out from the garbage. Further the presence of URL identified using a regular expression and remove all the URL's from the tweet. Remove all the private usernames identified by @user and the hash tags identified by the # symbol. The sample preprocessed data is given in table 2. To get the meaning of each word English dictionaries WordNet is used. The words that are not found illustrate that they are either slangs or abbreviations. The sample of the abbreviation list is given in table 4, SMS dictionary are used for this purpose. Further lemmatization applied to steam the word and apply corrections. Further the spell checking of the tweet is done in order to correct the effects of the lemmatizer. This step feeds the remaining words in the spell checker and substitute with the best match. To identify and remove the stop words we used the Stanford dictionary, which are then simply stripped from the tweet under process. Lastly we removed all the special characters excluding the emoticons. Only English tweets are considered for classification and analysis. The sample of resulted tweet is given in table 3.

Thus we normalized complete tweets by preprocess raw tweets and complete pre processed to produce meaningful tweets. Further completely processed the tweets data will be used for further opinion mining.

IV. CONCLUSION

Here, we developed the user friendly indigenous application based on windows system in Java with Netbeans IDE platform to extract, process and classify the real time tweets on the social keyword from the twitter web site, news website etc. using the twitter 4j Libraries and their API's. The extracted real time tweets are pre-processed for the removal of the URL, Hash tag using the Java regular expression. Further the pre-processed tweets were used for analysing abbreviation, slangs and removal of stop words to obtain meaning full tweets for the sentimental analysis.

ACKNOLEDGEMENT

I would like to specially thank to Mr. Ramanand Potdar and Mr. Amar Kokare for their valuable guidance and help during this work.

© 2015, IJCSE All Rights Reserved

TABLE 1 The sample of extracted real time and raw tweets on the 'Narendra Modi' [Mon Jun 16 20:52:20 IST 2014]

[Mon Jun 16 20:52:20 IST 2014]GET https://api.twitter.com/1.1/search/tweets.json?q=Narendra%20Modi&count=100&with_twitter_user_id=true&include_entities=true[Mon Jun 16 20:52:20 IST 2014]OAuth base string: GET&https%3A%2F%2Fapi.twitter.com%2F1.1%2Fsearch%2Ftweets.json&count%3D100%26include_entities%3Dtrue%26oauth_consumer_key%3Dux9athB6dnJLJlBQRLwe14Vkj%26oauth_nonce%3D51477371%26oauth_signature_method%3DHMAC-SHA1%26oauth_timestamp%3D1402932140%26oauth_token%3D163172408-ISIe3zS49ExyUSADxzgG5lWaPJkf8g9mIo0EPokd%26oauth_version%3D1.0%26q%3DNarendra%2520Modi%26with_twitter_user_id%3DtrueRetweets.","verified":false,"contributors_enabled":false,"profile_sidebar_border_color":"F2E195","name":"radhika r","profile_background_color":"BADFCD","created_at":"Wed Dec 30 05:07:03 +0000 2009","is_translation_enabled":false,"default_profile_image":false,"followers_count":216,"profile_image_url_https":"https://pbs.twimg.com/profile_images/430004058800140288/M1oTZ0os_normal.jpeg","geo_enabled":true,"profile_background_image_url":"http://abs.twimg.com/images/themes/theme12/bg.gif","profile_background_image_url_https":"https://abs.twimg.com/images/themes/theme12/bg.gif","follow_request_sent":false,"entities":{"description":{"urls":[]}},"url":null,"utc_offset":8000,"time_zone":"Quito","notifications":false,"profile_use_background_image":true,"friends_count":107,"profile_sidebar_fill_color":"FFF7CC","screen_name":"lotusprings","id_str":"100410807","profile_image_url":"http://pbs.twimg.com/profile_images/430004058800140288/M1oTZ0os_normal.jpeg","listed_count":0,"is_translator":false}},{"contributors":null,"text":"Top story:RT @toi Online posts against Narendra Modi spell trouble - The Times \u2026 http://t.co/qHJ3wLVGjW, see more http://t.co/8nYAAJq3cx","geo":null,"retweeted":false,"in_reply_to_screen_name":null,"possibly_sensitive":false,"truncated":false,"lang":"en","entities":{"symbols":[],"urls":[{"expanded_url":"http://timesofindia.indiatimes.com/india/Online-posts-against-Narendra-Modi-spell-trouble/articleshow/36635876.cms","indices":[81,103],"display_url":"timesofindia.indiatimes.com/india/Online-p\u2026","url":"http://t.co/qHJ3wLVGjW"},{"expanded_url":"http://tweetedtimes.com/search/the%20india%20times/en","indices":[114,136],"display_url":"tweetedtimes.com/search/the%20i\u2026","url":"http://t.co/8nYAAJq3cx"}],"hashtags":[],"user_mentions":[{"id":122587382,"name":"The Times of India","indices":[13,17],"screen_name":"toi","id_str":"122587382"}]},"in_reply_to_status_id_str":null,"id":478557996013060096,"source":"<a href=\"http://www.tweetedtimes.com\" rel=\"nofollow\">The Tweeted Times<\/a>","in_reply_to_user_id_str":null,"favorited":false,"in_reply_to_status_id":null,"retweet_count":0,"created_at":"Mon Jun 16 15:21:32 +0000 2014","in_reply_to_user_id":null,"favorite_count":0,"id_str":"478557996013060096","place":null,"user":{"location":"Nepal","default_profile":false,"profile_background_tile":false,"statuses_count":2476,"lang":"en","profile_link_color":"038543","profile_banner_url":"https://pbs.twimg.com/profile_banners/1903051339/1380091580","id":1903051339,"following":false,"protected":false,"favourites_count":1,"profile_text_color":"333333","description":"19, Failure is the first stair to success.#Football ||Don't follow your dreams; chase them.#Planets","verified":false,"contributors_enabled":false,"profile_sidebar_border_color":"EEEEEE","name":"Biplav Poudel","profile_background_color":"ACDED6","created_at":"Wed Sep 25 05:59:59 +0000 2013","is_translation_enabled":false,"default_profile_image":false,"followers_count":61,"profile_image_url_https":"https://pbs.twimg.com/profile_images/378800000505435829/746c6f9b0d09598c2370d7e519ff48d5_normal.jpeg","geo_enabled":false,"profile_background_image_url":"http://abs.twimg.com/images/themes/theme18/bg.gif","profile_background_image_url_https":"https://abs.twimg.com/images/themes/theme18/bg.gif","follow_request_sent":false,"entities":{"description":{"urls":[]}},"url":null,"utc_offset":null,"time_zone":null,"notifications":false,"profile_use_background_image":true,"friends_count":165,"profile_sidebar_fill_color":"F6F6F6","screen_name":"bpoudel01","id_str":"1903051339","profile_image_url":"http://pbs.twimg.com/profile_images/378800000505435829/746c6f9b0d09598c2370d7e519ff48d5_normal.jpeg","listed_count":12,"is_translator":false},"coordinates":null,"metadata":{"result_type":"recent","iso_language_code":"en"}},{"contributors":null,"text":"PM Modi Invited To Watch FIFA World Cup Final in Brazil http://t.co/3gNuYZje9v","geo":null,"retweeted":false,"in_reply_to_screen_name":null,"possibly_sensitive":false,"truncated":false,"lang":"en","entities":{"symbols":[],"urls":[{"expanded_url":"http://www.ndtv.com/article/india/pm-narendra-modi-invited-to-watch-fifa-world-cup-final-in-brazil-542433","indices":[56,78],"display_url":"ndtv.com/article/india/\u2026","url":"http://t.co/3gNuYZje9v"}],"hashtags":[],"user_mentions":[]},"in_reply_to_status_id_str":null,"id":478557975519313922,"source":"<a href=\"http://www.apple.com\"

© 2015, IJCSE All Rights Reserved

TABLE 2 The sample of preprocessed tweets related to the Search keyword 'Narendra Modi'

shared a link http://t.co/KlEibeXSlm ====
RT @RMantri: Huge - Modi govt to reform Industrial Disputes Act http://t.co/UhyPTTfGEI ‖ Failed and frustrated "experts" had said Modi won… ====
Top story:RT @toi Online posts against Narendra Modi spell trouble - The Times … http://t.co/qHJ3wLVGjW, see more http://t.co/8nYAAJq3cx ====
PM Modi Invited To Watch FIFA World Cup Final in Brazil http://t.co/3gNuYZje9v ====
@narendramodi please Mr.narendra modi bring one direction to India please we want one direction in India thank you -frm Indian directioners ====
RT @MrMikeStreet: Social Media applauds Shri Narendra Modi's message of India-Bhutan Friendship http://t.co/cbeqkh1dfu #in ====
RT @narendramodi: Here are some pictures of a memorable welcome I received in Bhutan http://t.co/GFh5TCVuIO ====
Modi, Sushma slip ups draw social media attention: Even as Prime Minister Narendra Modi and the Indian delegation... http://t.co/ih5EYamLal ====
RT @shilpitewari: No they just kill for supporting Narendra Modi RT @EchoOfIndia At least no one used to be arrested for criticizing Manmoh… ====
RT @ndtv: Narendra Modi's 'Nepal' slip in Bhutan fires up Twitter - NDTV http://t.co/2moMZfPYwH ====
shared a link http://t.co/CfTKvHM211 ====
RT @Narendramodi_G: Bhutan makes an exception for Narendra Modi, claps for him - http://t.co/GUSTGx5DWP ====
Social Media applauds Shri Narendra Modi's message of India-Bhutan Friendship http://t.co/cbeqkh1dfu #in ====
RT @NitiCentral: Narendra Modi seems determined to not repeat the mistakes of 1998-2004, bats for Digital BJP ǀ Writes @shashidigital http:… ====
RT @yesaashish: पीएम मोदी ने भूटान को कहा नेपाल, ट्विटर पर मचा बवाल http://t.co/1vwiGc12Su via @NavbharatTimes @NDTVravish #Feku #Modi ====
RT @bhaiyyajispeaks: Michael Schumacher just changed his life mode from Yogendra Yadav to Narendra Modi ====
RT @ProSyn: New from @ShashiTharoor: "Nehru's Last Stand?" http://t.co/rHu0ABxErf ====
PM Narendra Modi invited by Brazilian President to watch FIFA World cup finals - http://t.co/C2SFwvrzDO http://t.co/S6LK6HfMWD ====
Brazilian President invites Narendra Modi to Watch World Cup http://t.co/XCfqhZl8vc ====
Brazilian President invites Narendra Modi to Watch World Cup http://t.co/IPyfZnBxHK ====
Brazilian President invites Narendra Modi to Watch World Cup http://t.co/0rLZcU5EYT ====
RT @one_by_two: Whats the license plate # ?? #CokeCan RT@ndtv: PM Narendra Modi Travelled in his Armoured BMW in Thimphu ====

TABLE 3 The sample of completely processed tweets sample on the Narendra Modi

Whats the license plate ?? PM Narendra Modi Travelled in his Armored BMW in Thimphu
======== Huge - Modi govt to reform Industrial Disputes Act ‖ Failed and frustrated had said Modi
======== Delhi: PM Narendra Modi meets Defence Minister Arun Jaitely, Army Chief Gen. Bikram Singh and NSA in the war room of Indian
======== Watch Video: PM Shri Narendra Modi's address to Joint Session of the Parliament of Bhutan- Part4:
======== In defense of Narendra Modi government's plan to saffronise education in India
======== Watch Video: PM Shri Narendra Modi's address to Joint Session of the Parliament of Bhutan- Part 3:
======== PM concludes Bhutan tour
======== Narendra Modi's 'Nepal' slip in Bhutan fires up Twitter - NDTV
======== outdid his critics' foul imaginations by not going after bigwigs
======== New from Last Stand?"
======== Why Amrit Prajapati entered Narendra Modi's sabha with gun?
======== ministers have UPA-favoured babus as special staff
======== please share some "Confidential" info. on new Narendra Modi's Govt., please Expose Modi as soon as possible.JAI HIND
======== ISBPL: ministers have UPA-favoured babus as special staff: ministers have UPA-...

© 2015, IJCSE All Rights Reserved

======== : ministers have UPA-favoured babus as special staff: ministers hav...
======== lulz. Kitnon ko arrest karoge? Nine students in Kerala arrested for publishing anti-Narendra Modi puzzle in college
======== Narendra Modi
======== Nine students in Kerala arrested for publishing anti-Narendra Modi puzzle in college magazine
======== Here are some pictures of a memorable welcome I received in Bhutan
======== INDIAN PM Narendra Modi's speech acknowledging the Royal Family of Bhutan said, "I want to begin with a tribute...

TABLE 4. The sample abbreviations used of the for the conversion in to text

a&f#always and forever
a'ight#alright
a.i.m.#aol instant messanger
a/m#away message
a1t#anyone there
a3#anyplace, anywhere, anytime
aabf#as a best friend
aaf#as a friend
aak#Alive and Kicking
aamof#as a matter of fact
aboot#about
abreev#abbreviation
absnt#absent
abt#about

REFERENCES

[1]. Farhan Hassan Khan et. al. , TOM: Twitter opinion mining framework using hybrid classification scheme, Decision Support Systems, 2013, http://dx.doi.org/

[2]. 10.1016/j.dss.2013.09.004

[3]. A. Bifet, E. Frank, Sentiment Knowledge Discovery in Twitter Streaming Data,, Springer-Verlag, Berlin, Heidelberg, 2010, pp. 1–15.

[4]. A. Bifet, G. Holmes, B. Pfahringer, MOA-Tweet Reader: real-time analysis in twitter, streaming data, in: T. Elomaa, J. Hollm'en, H. Mannila (Eds.), DS 2011, LNCS 6926, Springer-Verlag, Berlin Heidelberg, 2011, pp. 46–60.

[5]. Argamon S., Bloom K., Esuli A., Sebastiani F., Automatically determining attitude type and force for sentiment analysis, in: Vetulani Z., Uszkoreit H. (Eds.), LTC2007, LNAI 5603, Springer-Verlag, Berlin Heidelberg, 2009, pp. 218–231.

[6]. Fu X., Guo Y., Guo W., Wang Z., et al., Aspect and sentiment extraction based oninformation-theoretic co-clustering, in: Wang J., Yen G.G., Polycarpou M.M.(Eds.),ISNN 2012, Part II, LNCS 7368, Springer-Verlag, Berlin, Heidelberg, 2012, pp. 326–335.

[7]. Lambodar Jena, Narendra Kumar Kamila, Data Extraction and Web page Categorization using Text Mining, *IJAIEM,* ISSN 2319 – 4847, Volume 2, Issue 6, June 2013

[8]. https://api.twitter.com/

[9]. https://dev.twitter.com/

[10]. http://search.twitter.com/search.format

[11]. http://ravikiranj.net/drupal/201205/code/machine-learning/how-build-twitter-sentiment-analyzer

[12]. J. Pasternack and D. Roth. Extracting article text from the web with maximum subsequence segmentation. In WWW '09: Proceedings of the 18th international conference on World Wide Web, page 1971{980, New York, NY, USA, 2010. ACM

[13]. Georgios Petasis1,2, Dimitrios Petasis1. BlogBuster: A tool for extracting corpora from the blogosphere. Software and Knowledge Engineering Laboratory National Centre for Scientific Research (N.C.S.R.) "Demokritos", Athens, Greece, 2010.

[14]. Baroni, M., Chantree, F., Kilgarri, A., Sharo, S. (2008). Cleaneval: a competition for cleaning web pages. In Proceedings of the 4th Web as Corpus Workshop (WAC4), - Can we beat Google?. N. Calzolari, K. Choukri, B.Maegaard, J. Mariani, J. Odjik, S. Piperidis, and D. Tapias, editors, Proceedings of the 6th International Language Resources and Evaluation (LREC 2008). Marrakech, Morocco, 2008

[15]. S. Evert. A lightweight and ecient tool for cleaning web pages. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis, and D. Tapias, editors, Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), Marrakech, Morocco, may 2008. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/N.B. Salem, and J-P Hubaux, "Securing Wireless Mesh Networks", IEEE Wireless Communications, Vol.13, Issue-2, **2006**, pp.**50-55**.

[16]. S. Han, E. Chang, L. Gao, T. Dillon, T., Taxonomy of Attacks on Wireless Sensor Networks, in the Proceedings of the 1st European Conference on Computer Network Defence (EC2ND), University of Glamorgan, UK, Springer Press, SpringerLink Date: December **2007.**

[17]. C. Karlof and D. Wagner, "Secure routing in wireless sensor networks: attacks and countermeasures," Ad Hoc Networks 1, **2003**, pp. **293-315**.

© 2015, IJCSE All Rights Reserved

[18]. Y. Yang, Y. Gu, X. Tan and L. Ma, "A New Wireless Mesh Networks Authentication Scheme Based on Threshold Method," 9th International Conference for Young Computer Scientists (ICYCS-2008), **2008**, pp. **2260-2265**

AUTHORS PROFILE

Bapurao M Bandgar received his MCA Degree in 2010, M.Sc. (Physics) degree in 1999 and B.Sc. (physics) in 1995 from Pune University. Perusing Ph.D(Comp. Sci.) from Karpagam University. His research interest includes Social Network Analysis and Data Mining, Sentimental Analysis.

Dr. Binod Kumar is Director & Professor at JSPM's Jayawant Technical Campus, affiliated to University of Pune , India . He has been in the field of teaching since more than 16 years. At present six Ph.D. candidates are registered under his supervision. He has published nearly 30 papers in International and National Journals. His areas of interest are Data Mining, Bioinformatics and Software Engineering

© 2015, IJCSE All Rights Reserved