Automating Corpora Generation with Semantic Cleaning and Tagging of Tweets for Multi-dimensional Social Media Analytics

Nazura Javed
Research Scholar
Bangalore University
Bangalore, India

Muralidhara B.L.
Bangalore University
Bangalore, India

ABSTRACT
Developing corpora from social media content involves convoluted cleaning. In this paper we propose and implement the automation of corpora building for facilitating social media mining and analytics. This automation process incorporates: a) metadata extraction and structuring b) semantic cleaning with tagging and c) learning domain terms/entities. The implementation performs comprehensive cleaning including abbreviation and slang correction, phonetic matching using metaphone algorithm, splitting joined words and identifying/learning entities. It identifies the entities, tags them and creates/updates a knowledgebase (KB) comprising of domain terms. The corpus thus constructed, facilitates multidimensional analysis and summarization. This proposed technique was tested with an experiment in which real world streaming tweets pertaining to Indian politics were collected, structured, cleaned and tagged. The results of the automation experiment can be stated as follows: a) the tweets although primarily in English, contained at times words from the regional languages. The algorithm does not recognize these words and they are construed as domain terms. An accuracy of 85.55% was achieved in identifying the correct domain terms and entities. b) The automation required human feedback and intervention which progressively reduced and reached a figure of 18% with the update and enhancement of the KB. This paper assumes relevance because the implementation *automates* the entire process of collecting and cleaning the tweets and yields a corpus suitable for multi-faceted analysis.

General Terms
Corpus Generation, Social Media, Multi-dimensional Analytics, Text Mining

Keywords
Corpora, Tweets, Social Media, Mining, Analytics, Knowledgebase

1. INTRODUCTION
Social media mining and analytics have assumed relevance today because of the popularity and inclusiveness of social media, increasing volume of content and the consequent need for gaining insight into the diffusion of information, opinions, sentiments and trends. Twitter is a micro blogging service that allows people to communicate with short 140 character messages. It is increasingly used by society to express views, sentiments, concerns and debate issues. It has gained significance in the political context with political leaders and entities leveraging it to disseminate, promote, support and debate causes. Extracting, tagging, classifying and summarizing the tweets is necessary so as to provide an insight into societal trends and scenarios. However mining of tweets can be effectively done only if a clean corpus is available. Cleaning of tweets becomes a challenging task because of the following reasons:

a. Unstructured tweet text with non adherence to the grammatical syntax.

b. Incorrect spelling and words spelled according to the phonetics.

c. Non English content i.e. use of regional language along with English.

d. Use of slangs and non standard abbreviations.

e. Multiple words combined together for e.g. profarmercongress – which require splitting into pro, farmers, Congress. In this paper we refer to them as joined words.

f. Use of numerals in text for e.g. gr8 for great, 2 good for too good.

In this paper, we focus on the following: a) Automatic corpus generation with the objective of reducing the efforts involved in capturing, cleaning, pre-processing and thus facilitating analytics. b) Generating a corpus which is amenable for Named Entity Recognition (NER), multidimensional analysis and view of the domain and c) Building and enhancing a domain KB. We also perform metadata extraction, structuring and semantic tagging in the course of automation. This helps us to avoid the use of additional processing resources for separate metadata handling and provides scope for multidimensional analysis by relating the text with the metadata.

The experiment conducted by us extracts #tagged and @tagged words as potential topics or entities, structures the metadata such as location, retweet count, followers count so as to gauge the regional preferences and popularity, performs comprehensive cleaning, and builds/enhances a KB comprising of political terms and entities.

As per our study and to the best of our knowledge, there is very little literature which addresses automatic cleaning techniques for real time tweets while also incorporating machine learning and facilitating a multidimensional political view.

The remainder of the paper is structured as follows: In section 2 we overview the related literature. Section 3 proposes the methodology for corpora building and tagging. Section 4 describes the experiments and the results thereof. This paper concludes by examining the scope/ boundary of the proposed

techniques and future scope of enhancement and research.

2. RELATED WORK

Social media analytics involves a three-stage process: *capture*, *understand*, and *present*. The *capture* stage involves obtaining relevant social media data by monitoring or "listening" to various social media sources, archiving relevant data and extracting pertinent information. All the data extracted is not useful and hence the *understand* stage selects relevant data for modeling, removes noisy, low quality data, and employs various advanced data analytic methods to analyze the data retained and gain insights from it [1]. Social media analytics is concerned with developing and evaluating informatics tools and frameworks to collect, monitor, analyze, summarize, and visualize social media data, usually driven by specific requirements from a target application. However, social media analytics faces several challenges such as semantic inconsistency/inaccuracies, misinformation and lack of structure as well as dynamic nature of social media data and their sheer size [2]. The limited length of a tweet and no restrictions on its writing styles results in grammatical errors, misspellings, and informal abbreviations [3]. Mining social media poses challenges because of the use of abbreviations, phonetic substitutions and structure. It is necessary to normalize and convert into standard format [4]. Works [2, 3, 4] thus highlight the relevance of structuring and cleaning the tweets for reliable semantic analytics.

Twitter analysis was done with archived dataset in [5]. Cleaning, sanitizing, parsing and storing the tweet text, and also details such as the number of retweets, replies and hashtags were done. Statistical summaries regarding retweets were computed; letter comparisons were made to uncover changes. Hashtag analysis was done to extract topics. Our work too, addresses the above issues but however, since our dataset is comprised of real time tweets, we subject them to a more comprehensive cleaning procedure. Word cleaning algorithm was proposed [6]. This algorithm removed non-English tweets, dropped retweets, removed URLs, hashtags and tackled the problems of repetitive letters. The significance of cleaning and preprocessing is highlighted [7]. This paper proposed a framework for preparing and using corpora from Social networks .The framework comprised of three phases. The first phase cleaned and preprocessed the data collected. The second phase applied various text processing techniques on the prepared corpora and the third phase performed text classification. Content was cleaned by removing URLs, removing non English words, removing spelling mistakes, and replacing abbreviations.

Data selection and filtering are usually based on keywords such as named entities or metadata released by micro-blog authors. Metadata on time and geolocations; the user's age, gender, background, and social environment enable the detection of sentiment variations or trends in the different groups and regions. The relevance of metadata data is highlighted in [8]. [9, 10] discuss the relevance of metadata and its applications. Extracting facets from the tweets in order to facilitate multi-faceted *search* was explored [9]. The entities identified were tagged to facilitate search. In our implementation, we too extract the metadata and entities in the course of cleaning. But our objective for extraction is to facilitate *analysis* of multidimensional nature. Adding Twitter specific metadata to the social network graph, can lend rich expressiveness for later analysis. Node information is enhanced with the user name, the status count and friends/followers count whenever the tweets are parsed by the social network analyzer [10]. Information credibility of a Twitter tweet can be determined on the basis of the presence of URL, the number of retweets and the length of the tweets [11]. In our, work, we capture the real time tweets as well as metadata so as enable us to relate the tweet text with the metadata and perform a multi-faceted analysis.

The contribution of our paper is as follows:

a) Automation of the entire process of collection, filtering, structuring, storing, cleaning and enhancing of tweets so as to derive a suitable corpus. b) Comprehensive cleaning with the help of an English lexicon; using slang and SMS dictionary for handling slang/abbreviations and replacing them; use of Metaphone algorithm for finding phonetic matches for incorrectly spelled words and splitting joined words using a English lexicon, domain KB and dynamic programming algorithm c) Building/enhancing the domain KB to recognize and enhance domain terms. d) Use of metadata for multi-faceted view and analysis.

3. METHODOLOGY

Twitter is a micro blogging service that allows people to communicate with short 140 character messages. Twitter has become a platform for information dissemination and change. Most of the governments, ministries, institutions and activists are on the platform. Twitter has emerged as a free source of breaking news and a force influencing the discourse in the country [12]. Twitter provides an open platform for expression, and hence mining of tweets can provide valuable insight into political scenario and societal opinion. Tweets can be collected using the REST or Streaming API. Although the tweet text is limited to just 140 characters, there is an associated rich metadata like the location, time zone, geographical latitude, geographical longitude, creation date, followers count and retweet count that can be captured and utilized for analysis. The retweet count is an indicator of the popularity of the tweet. The geographical parameters and location indicate the geographical origin of the tweet. Mining the text in conjunction with metadata can provide an insight into the different political perspectives, political trends, sentiments and opinions. However, the effectiveness of mining tweets is based on the availability of a clean corpus.

Our study indicates that real world tweets require a comprehensive procedure for removing noise. Apart from customary cleaning, like removing duplicates, repeating characters, special characters, character encodings, we address the follows issues.

a) Non standard abbreviations and slangs: The language used in the tweet is not just restricted to the vocabulary of the slang/ SMS directory. Non standard abbreviations like for e.g. 'nw' for 'now', 'thr' for 'there' are used. The Metaphone algorithm incorporated in our implementation performs phonetic matching helps at arriving at the correct suggestions. Our implementation first checks the existing slang directory for the known slangs. The known slangs are replaced by their correct full forms. The unknown or non standard slangs are phonetically matched to derive a list of suggestions. An array of non standard abbreviations is then built /concatenated to contain these non standard abbreviations with their appropriate correct forms.

b) Joined words in tweets: Since the twitter text is restricted to 140 characters, users generally try to squeeze in more content by joining multiple words. For e.g. the joined word 'continueconversionagenda' requires to be split into the words continue, conversion and agenda. The

joined word is flagged as misspelled, when a lexicon matching/search is carried out. The splitting technique used by us, makes use of enchant spelling library, the domain KB and dynamic programming algorithm to arrive at the optimum split suggestions. Use of the Domain KB at the time of splitting enables splitting of joined words containing domain entities.

c) Use of Regional terms: Though the tweets collected were primarily in English, they contained at times words from the regional languages. These words though spelled using English alphabets are not a part of English lexicon for e.g. 'Pradhan Mantri' which means 'Prime Minister' in Hindi language. Our algorithm tags them as domain terms and updates the domain KB. Though this approach is not essentially correct, it has the advantage that, these words are not flagged as *misspelled* while cleaning and processing later tweets. This also provides scope of building *a lexicon of commonly used regional words spelt in English* for future analysis.

Figure 1 provides an overview of the methodology followed. The proposed methodology implements the automation of collection, cleaning and tagging of tweets. The scope of this technique is limited to building corpora which is acquiescent for mining. The corpus so developed has a potential of being used for a multi-faceted view of the social media content.

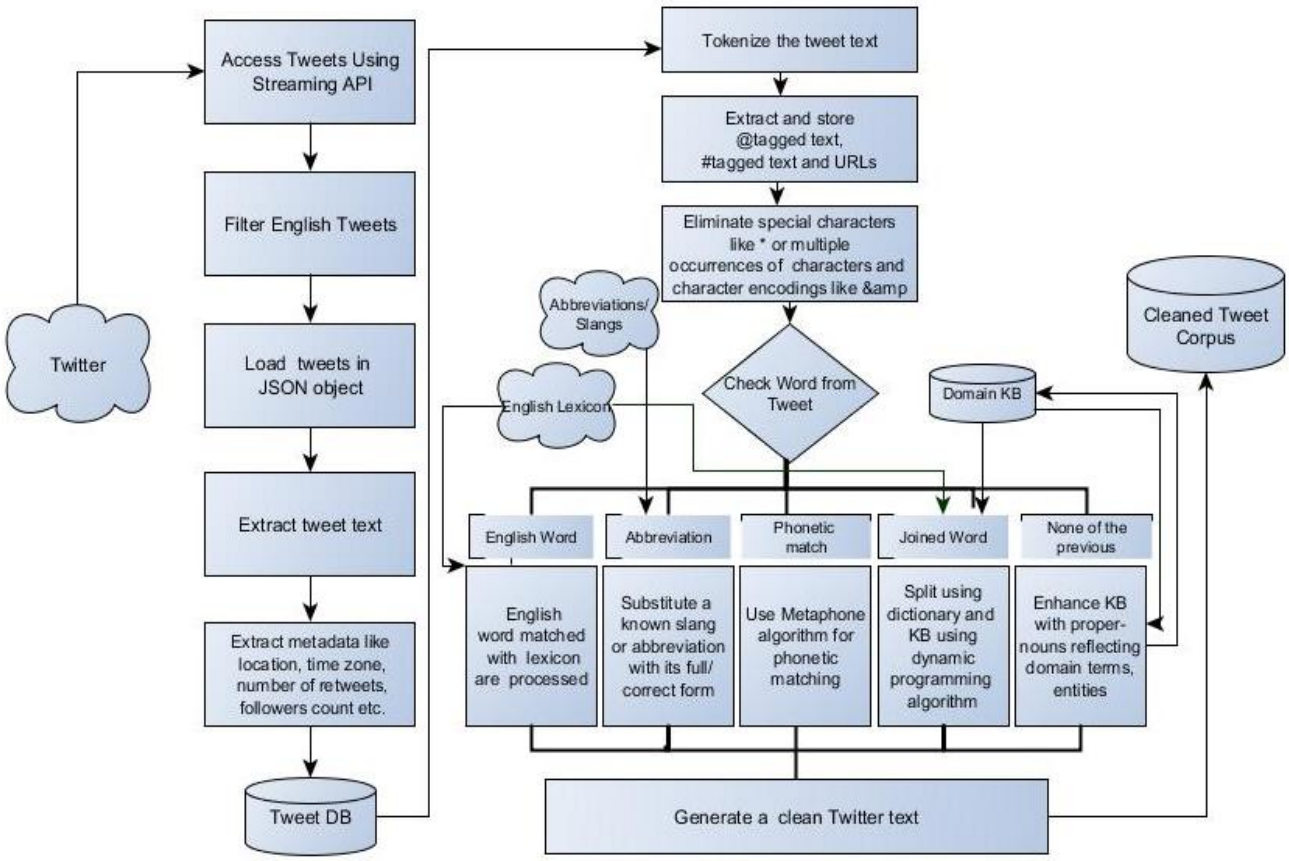

Figure 1: Methodology: Automation of Twitter Corpus Generation

3.1 Algorithm 'processTweet'

We propose the *'processTweet'* algorithm which collects, filters, extracts, structures, cleans and stores tweets for a later multidimensional political analysis. It simultaneously learns the domain terms and builds/enhances a Domain KB. It extracts and structures the metadata with an objective to: a) discover relationship between the political sentiments and geographical location, b) obtain date wise important events, topics of discussions and controversies with supporting statistics and c) determine the popular tweets and the most discussed and debated issues.

Table 1. Algorithm 'processTweet' Terms and Terminology

t_n	Tweets 1 to n
w_{im}	Tokens 1 to m in each tweet i
TweetDB	Database containing real world tweets and metadata
AbbreviationLexicon	A standard slang lexicon containing popular slangs and abbreviations
AbbreviationTable	Vector updated with non standard abbreviations and their full forms
EnchantLexicon	The Python Lexicon
DomainKB	KB updated with domain terms
TwitterCorpus	Corpus comprising of cleaned tweets, domain terms and tagged terms

***Algorithm processTweet*(Tweets, AbbreviationLexicon, AbbreviationTable, EnchantLexicon, DomainKB, TweetDB, TwitterCorpus)**

Input : Tweets, Abbreviation Lexicon, AbbreviationTable, EnchantLexicon, DomainKB
Output: *A* TwitterCorpus *database containing the cleaned tweets*; *an updated* DomainKB

```
set OAuth setting ;
set criteria for streaming ;
capture the streaming Tweets;
remove duplicate Tweets;
load Tweets in JSON object;
filter English Tweets;
Extract metadata: username, location, timezone, geolat, geolon, retweet count, follower count;
update TweetDB with Tweet text and metadata;
set t_n = read Tweets from TweetDB;
for each Tweet i in t_n
        set processedtweet_i = null;
        set w_im = tokenize Tweet t_i into m tokens;
        for each token j in w_im
                if w_ij contains '@' or w_ij contains '#' or w_ij contains 'http'
                        extract w_ij;
                        processedtweet_i = processedtweet_i ∪ (w_ij − '@') ∪ (w_ij − '#');
                end if
                if w_ij ∈ special characters or w_ij ∈ character encodings
                        skip w_ij;
                end if
                if w_ij ∈ EnchantLexicon
                        processedtweet_i = processedtweet_i ∪ w_ij;
                else
                        if w_ij ∈ AbbreviationLexicon
                                processedtweet_i = processedtweet_i ∪ expandedword;
                        else if w_ij ∈ DomainKB
                                processedtweet_i = processedtweet_i ∪ w_ij;
                        else if w_ij has metaphone match
                                processedtweet_i = processedtweet_i ∪ matched_word;
                                AbbreviationTable = AbbreviationTable ∪ (w_ij, matched_word);
                        else if w_ij is a joined word
                                split_w_ij = spilt using EnchantLexicon and DomainKB;
                                processedtweet_i = processedtweet_i ∪ split_w_ij;
                        else
                                DomainKB = DomainKB ∪ w_ij;
                                processedtweet_i = processedtweet_i ∪ w_ij;
                end if
        end for;
        update database TwitterCorpus with processedtweet_i, @tagged_data_i,
                #tagged_data_i and domain_terms_i;
end for;
end processTweet;
```

4. EXPERIMENT AND RESULTS

The 'processTweet' algorithm which automates the process of collecting and cleaning the tweets was coded using Python 2.7. 8220 tweets were collected using the Python Tweepy library over a time window, for a period of 3 days between the 28th and 30th June. Unprocessed as well as the cleaned, processed tweets were stored in MySql database. PyEnchant a spellchecking library for Python was used for checking the spelling of English words and for splitting the joined words. Jellyfish, a Python library was used for doing approximate and phonetic matching using the metaphone algorithm. A customized slang/abbreviation dictionary was used for replacing the standard slangs and abbreviations. The experiment was conducted by filtering English tweets relating to Indian politics and Prime Minister 'Narendra Modi'.

4.1 Cleaning

The python code transforms a tweet into a clean or processed tweet. Figure 2 demonstrates the execution of the automated cleaning program. The automation process requires human intervention/feedback at the time of: a) replacing the new slangs b) performing splitting c) extracting proper nouns representing domain terms and entities. The human intervention required in the initial iterations progressively reduces to approximately 18% as the automation code learns the domain terms and slangs.

```
@imjadeja   ; @imraina  guys  !  !  what do you want to say about lalit modi's email ?
we trust u  !  please come ahead  ; give ans to this rubbish.
Original:   @imjadeja   ; @imraina  guys  !  !  what do you want to say about lalit modi's email ?
we trust u  !  please come ahead  ; give ans to this rubbish.
Cleaned:     imjadeja ;imraina guys !!what do you want to say about Lalit Modis email ?we trust you !please come ahead ;give answer to thi
s rubbish
Entities & Terms: Lalit Modis
@denniscricket_  icc still believes lalit modi after his failed accussation against nz allrounder
accussation To accusation
Accept Above Correction Y/N y
nz To NZ
Accept Above Correction Y/N y
allrounder To all rounder
Accept this Above Correction Y/N y
Original:   @denniscricket_  icc still believes lalit modi after his failed accussation against nz allrounder
Cleaned:    denniscricket_ Icc still believes Lalit Modi after his failed accusation against NZ all rounder
Entities & Terms: Icc Lalit Modi
```

Figure 2: The Execution of the 'processTweet' process

4.2 Metadata Extraction and Structuring

The metadata like @tags, #tags, username, geographical location, statistical metadata like the followers count and retweet count can be leveraged for a later multi-faceted analysis. These metadata are extracted, organized and stored in the MySql database along with the processed or cleaned tweet. Table 2 shows an extract of the captured tweets, processed tweets along with the associated metadata and tags like location and domain terms.

Table 2. An extract of the Captured and processed tweets with Metadata

Tweet	location	Processed Tweet	tag_at(@)	Domain terms
rt @jhasanjay: mr modi publicly confesses that his is a suit-boot ki sarkar ! but who carries your suitcase, sir? a certain mr adani who glâ€¦	Siegen, Germany	retweet: jhasanjay : Mr. Modi publicly confesses that his is a Suitboot ki Sarkar ! but who carries your suitcase ,sir ? a certain Mr. Adani who gl	@jhasanjay	Modi Suitboot Ki Sarkar Adani gl
govt committed to roll out one rank, one pension scheme: pm narendra modi ... - financial express http://t.co/okqyw5fdnx	India	Government committed to roll out one rank ,one pension scheme :PM Narendra Modi - financial express		Narendra Modi
pm modiâ€™s bangladesh visit: teesta pact not on this trip, says sushma swaraj http://t.co/ixl1smpesy	London	PM Modis Bangladesh visit :Teesta pact not on this trip ,says Sushma Swaraj		Modi Bangladesh Teesta Sushma Swaraj

4.3 Observations

The results of the 'processTweet' implementation can be stated as under:

a) It successfully removes the duplicate/retweeted tweets, @tags , #tags , URLS, repeating characters, special characters and character encodings like &amp.

b) It replaces the known slangs with their full forms and also handles the unknown and non standard slangs by coming up with suggestions based on phonetic matches.

c) It arrives at the suggestions for splitting the joined words.

d) The words which are not recognized by the English Lexicon; the words which cannot be split or the slang words that cannot be expanded to their complete, correct forms get included as domain terms. These words are extracted and stored separately in the MySql database. The KB is updated and enhanced with these terms.

e) It was observed that i) regional terms, spelled in English for e.g. Sarkar get tagged as domain terms. ii) the words at the end of the tweet which get cut off because of 140 characters size limit of the tweet text, cannot be identified and are construed as keywords. Thus i) and ii) contribute to the False Positive results. Table 3 summarizes the accuracy of the domain term identification.

Table 3. Accuracy of Domain Term Identification

Description	Number	%
Correctly identified Domain	14025	85.55%

Terms/Entities (True Positives)		
Regional terms falsely construed as Domain Terms/Entities (False Positives)	1956	11.92%
Slangs/Abbreviations falsely construed as Domain Terms/Entities (False Positives)	416	2.53%
Total terms identified/inferred	16397	100.00%

5. CONCLUSION

Automation of the process of collection, cleaning and tagging of tweets reduces the manual effort in developing a corpus amenable for mining. This work assumes relevance because the implementation *automates* the entire process of collecting and cleaning the tweets and yields a corpus suitable for multi-faceted analysis. It enables the development of corpus which is suitable for analysis of different topics like political issues, sentiments of the society, product analysis, feedback analysis etc. Tagging the tweets using metadata and building a KB of domain terms lays a foundation for interpreting, analyzing and providing multidimensional views. In the course of the experiment, the following limitations were observed. a) the words at the end of the tweet which get cut off because of the 140 characters size limit of the tweet text, cannot be identified by the automation algorithm b) some of the tweets when cleaned, do not contain any meaningful content. Discovering these irrelevant tweets or determining the criteria for filtering them and automation of the same can help to refine the above process. Identifying these irrelevant tweets, eliminating them and using text mining techniques to interpret them would be the next milestone.

6. REFERENCES

[1] Fan, W., & Gordon, M. D. (2014). The power of social media analytics.Communications of the ACM, 57(6), 74-81.

[2] Zeng, D., Chen, H., Lusch, R., & Li, S. H. (2010). Social media analytics and intelligence. Intelligent Systems, IEEE, 25(6), 13-16.

[3] Li, C., Sun, A., Weng, J., & He, Q. (2015). Tweet Segmentation and its Application to Named Entity Recognition. Knowledge and Data Engineering, IEEE Transactions on, 27(2), 558-570.

[4] Han, B., & Baldwin, T. (2011, June). Lexical normalisation of short text messages: Makn sens a# twitter. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1 (pp. 368-378). Association for Computational Linguistics.

[5] Chen, B., Chen, X., & Xing, W. (2015, March). Twitter Archeology of learning analytics and knowledge conferences. In Proceedings of the Fifth International Conference on Learning Analytics And Knowledge (pp. 340-349). ACM.

[6] Xiang, G., Fan, B., Wang, L., Hong, J., & Rose, C. (2012, October). Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. In Proceedings of the 21st ACM international conference on Information and knowledge management (pp. 1980-1984). ACM.

[7] Medhat, W., Yousef, A. H., & Korashy, H. (2014, November). A Framework of preparing corpora from Social Network sites for Sentiment Analysis. InInformation Society (i-Society), 2014 International Conference on (pp. 32-39). IEEE.

[8] Bosco, C., Patti, V., & Bolioli, A. (2013). Developing corpora for sentiment analysis: The case of irony and senti-tut. IEEE Intelligent Systems, (2), 55-63.

[9] Abel, F., Celik, I., Houben, G. J., & Siehndel, P. (2011). Leveraging the semantics of tweets for adaptive faceted search on twitter. In The Semantic Web–ISWC 2011 (pp. 1-17). Springer Berlin Heidelberg.

[10] Klein, B., Laiseca, X., Casado-Mansilla, D., López-de-Ipiña, D., & Nespral, A. P. (2012). Detection and extracting of emergency knowledge from twitter streams. In Ubiquitous Computing and Ambient Intelligence (pp. 462-469). Springer Berlin Heidelberg.

[11] O'Donovan, J., Kang, B., Meyer, G., Hollerer, T., & Adalii, S. (2012, September). Credibility in context: An analysis of feature distributions in twitter. In Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom) (pp. 293-301). IEEE.

[12] Times of India 19th July 2015 "Governments have understood the potential of social media", Arun.Dev@timesgroup.com

[13] Zappavigna, M. (2012). Discourse of Twitter and social media: How we use language to create affiliation on the web. A&C Black.

[14] Russell, M. A. (2013). Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More. " O 'Reilly Media, Inc.".