International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-6, Issue-2 E-ISSN: 2347-2693

Event Extraction from Twitter using Scoring Function and LDA

Monika Gupta[1*], Parul Gupta[2]

[1*]Computer Science, Y.M.C.A. University of Science and Technology, Faridabad, India
[2]Computer Science, Y.M.C.A. University of Science and Technology, Faridabad, India

Corresponding Author: monika.mittal167@gmail.com, Tel: 8950602237

Available online at: www.ijcseonline.org

Received: 31/Jan//2018, Revised: 08/Feb2018, Accepted: 20/Feb/2018, Published: 28/Feb/2018

Abstract- Extracting and interpreting information from user generated content is a current topic in the scientific community and in the business world. Furthermore, data with a spatial component are even more important. This is proved by the numerous web applications that deal with processing and visualization of user generated content. The task of this extraction is to collect major life events in the form of retrievable entries that include structured data about major life event name, location and time which are often, categorized by complex, and nested structures involving ambiguous entities.

Keywords: Events, tweets, LDA, Extraction

I. INTRODUCTION

Huge collection of research work has focused on extracting event summary beyond merely discovering a set of entities representing the event. Events are also defined as statements that describe the circumstances in which something holds true. Events may be expressed by means of nominalizations, verbs, adjectives, Predicative clauses, or prepositional phrases. It is one of the atomic operations in detection and involvement among entities and other information in tweets or documents. This extraction is useful for the various professional news journalists; it helps them to apply social media content as an information source helps to get a handle on with the huge amount of information. Event extraction combines knowledge and experience from a number of domains; it includes computer science, linguistics, NLP, data mining, and artificial intelligence. With the advent of online media and the social web, monitoring online reputation has become important for any organization. Companies are keen on information spreading through word of mouth in social networks. Recently there has been a lot of work on event detection from Twitter. Example is shown in Figure 1. Tracking their reputation in social media is an important task for assessing overall sentiment across people about their products. Regular alerts can be sent, based on severity of the topics that are critical to company's reputation. Formally, the task of event extraction is to identify major events category in twitter data and to find detailed information about them, ideally identifying when, with what methods, and where. The vast use of social media to share facts and opinions about entities, such as current trends, brands, companies and figures has generated the necessity of managing the online reputation of those entities.

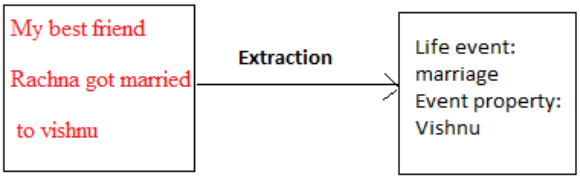

Figure 1. Example of Event Extraction

A summary of detected events is useful for a large number of applications like forest-fires tracking, sporting events detection, finding local festivals, detecting drug related adverse events, traffic events, epidemics, earthquakes, emerging controversial events, etc. Automatically extracting events is a higher-level information extraction task which is highly complex because of natural language processing and also due to the fact that in twitter data a full event description is usually scattered over several sentences and articles also more comprehensive.

II. OBJECTIVES

The aim of extracting events from a twitter page can be divided into several key Sub-problems:1. Identify if a twitter page contains events. 2. Find the list of events on the page. 3. Extract required information from these events. 4. Store information about an event in a database. The main focus within this research is on extraction of information from events which is depicted in sub problem 3. There are different sources which offer events and they all have a common goal,

to make the information easy to understand and use for the users. This is often extracted by having some sort of list with information about each event. The problem is that almost all the sources of events differ from each other in some way, i.e. in how the information is stored. The information is also generated by people, therefore the information often contain inconsistencies. This often comes from that there could be different persons that are writing the same information but also that the structure of the source is not built for handling information about Events.

III. RELATED WORK

A review of previous work on extraction of events from twitter is given and how these methods measure up to the requirements of the twitter data analysis and extraction has been discussed. Several researchers have been studying the twitter data analysis and numerous architectures appear in the literature.

Related work for the topic is as follows:

Mohammad AL-smadi and Omar Qawasmeh [1] proposed a knowledge –based approach for event extraction from Arabic tweets. The main objective of their paper was to extract the events from Arabic tweets by using knowledge based approach. John foley, Michael Benderky and Vanja Josifovski [2,13] proposed the method of local event extraction from the web. The main objective of their paper was to provide the scoring function on document, region and field-set. Jiwei Li, Alan Ritter, Claire Cardie and Eduard Hovy [3,16] proposed the Major Life Event Extraction from Twitter based on Congratulations/Condolences Speech Acts. They provide the method of LDA (topic modeling) and human screening approach for extracting the events from twitter. G. Katsios, S. Vakulenko , A. Krithara and G. Paliouras [4,10] proposed open domain Event extraction from twitter: Revealing Entity Relations. The main objective of their paper was to extract Events based on Named entity Recognition, Relation selection and Ranking Approach. Feifan Liu, Jinying Chen, Abhyuday Jagannathha, Hong [5] Yu proposed learning from Biomedical Information Extraction: Methodology Review of Recent Advances. Biomedical information extraction aims to automatically unlock structured semantics out of unstructured biomedical text data. Abdur Rahman M.A. Basher, Alexander S. Purdy and Inanc Birol [6,12,14] proposed Event Extraction from Biomedical Literature. Their work provide the opportunity to extract accurate context of the observed mutations to cancer and treatment, as well as the opportunity to generate new hypotheses by discovering and assessing novel relationships among entities in literature and genomic data. Frederik Hogenboom, Flavius Frasincar, Uzay Kaymak, Franciska de Jong and Emiel caron [7,17] proposed a survey of Event Extraction methods from text for decision support systems. Dr. D Ramesh, Dr.S.Suresh kumar [8] proposed the method of Event Extraction from Natural language Text .They proposed the framework which consists of 3 subtasks namely Preprocessor,POS Tagger and Event Extraction Modules. Kolikipogu Ramakrishna, Vanitha Guda, Dr.B.Padmaja Rani , Vinaya Ch [9,11,15] proposed a novel model for timed event extraction and temporal reasoning in legal text documents. They give the framework which consists of four subsystems 1. NLP system. 2. Annotation structure and tagger for temporal expressions and events. 3. Post processor including a knowledge-based sub system and 4. A reasoning mechanism which models temporal events in temporal constraint satisfaction networks (TCSNS).

IV. TOPIC MODELING

Recently Topic Modeling has gained attention for discovering topics in a large collection of documents. Topic modeling approaches like LDA, however takes only the co-occurrence information into account available from the data, but the semantic dependencies and the sub-topic structure are not handled by LDA. Categories are identified based on the words set in the table. There are some difficulties that are faced by to recognize the synonyms and same word event category. To recognize this task we need to identify each word and its meaning. This is actually very time consuming and more error prone. An LDA based topic modeling is used to group the collection of tweets to find major life event categories in an unsupervised way. Table 1 shows the event categories of some topics.

Table 1. Event Types Classification

Human Label	***Tag Words***
Wedding & Engagement	Wedding, love, ring, engagement ,engaged, bride, marriage
Admission	Admitted, university, college, offer, school, acceptance, profession
Exam	Exam, test, semester ,Exams ,practical, assessment, oral
Research	Research ,presentation, journalism, paper, conference, go, writing
Movie	House, movie ,city, home, place, town, leaving
Vacation	Vacation, family ,trip, country, go, flying

V. PIPELINED ARCHITECTURE OF THE PROPOSED SYSTEM

From Figure 2 Pipeline A first examines the event category of the input tweets speaks about and removes the irrelevant and unimportant tweets. Next, Pipeline B identifies whether the speaker is directly belongs to the event. Finally, Pipeline C extracts the event category. In this Pipeline A extracts the major life events tweets such as I got married to vishal and my friend rachna got married and remove the tweets such as I had rice for lunch. Pipeline B takes the output of first pipeline as input and identifies whether author is directly related or not and based on this filters the tweets such as My friend rachna got married.

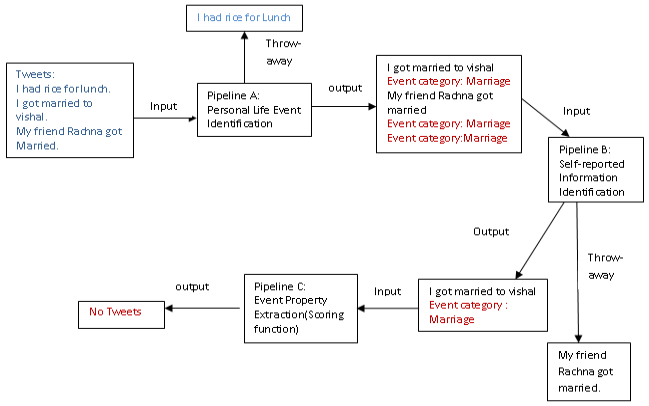

Figure 2. Pipelined Architecture of proposed algorithm

Finally, pipeline C takes the output of second pipeline as input this pipeline extracts the events from tweets.
Here, clearly shown that pipeline C contains no results Because past related tweets are not allowed in this pipelined architecture of the system. This method correctly filters past related tweets.

VI. PROPOSED ALGORITHM FOR EVENT EXTRACTION FROM TWITTER

An algorithm is proposed for Event Extraction from twitter.

Input: Reply list R = {f}, Tweet collection Tw= {l}, Retrieved Tweets database C = ϕ.
Identified topic list M=ϕ
Begin
While not stopping:
1. For unprocessed conversation l ϵ Tw if l contains reply f ϵ R,
- add l to C: C = C + l.
- remove l from Tw: Tw = Tw - l

2. Run LDA on newly added tweets in C.
3. Identify meaningful/unimportant topics, giving label to meaningful topics.
4. Add newly detected important topic l to M.
5. For conversation l belonging to unimportant topics
- remove l from C: C = C - l

6. Collect more tweets based on topic modeling.
7. Provide score ϕ(F) to each tweet.
- If ϕ(F)>1

8. Add meaningful responses to R.
End
Output: Identified topic list M. Tweet collection C.

Figure 3. Proposed algorithm of Event Extraction

In this, Human screening is completely eliminated by the algorithm so, because of this results are more accurate and consume less time also training is not as much required in earlier algorithm. Past related tweets are eliminated up to some extent not completely. Event collision is also removed by this approach. Event collision is occurring when tweet belongs to more than one event category. Tweets are collected and for unprocessed tweets run the algorithm as show in Figure 3. For unprocessed tweets run LDA streaming and retrieve the tweet based on the domain in which it belong. If the tweet belong to certain event category than extract tweets based on domain based classification. If tweet does not belong to any domain then discard that tweet. For meaningful tweet apply the scoring function ϕ(F) and test whether the tweet belong to event category. If it belongs to event category then we store in the database otherwise delete from the dataset.

VII. PROPOSED METHODOLOGY

To generate more accurate results jointly scoring function is used which effectively and efficiently categorizes the event. In this LDA and scoring function is used to correct identification of events. LDA is used to separate the domains. Useful tweet goes under their respective domain category other trash topic goes under another category known as unknown category. Figure 4. Shows the methodology of the proposed algorithm.

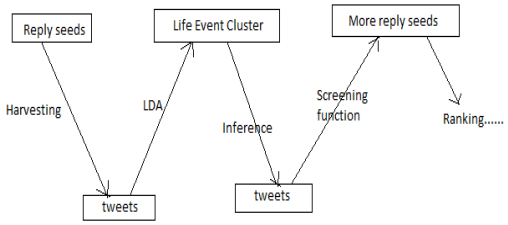

Figure 4. Proposed Methodology for Event Extraction

Scoring function is used to correctly identify whether the tweet is belonging to that domain or not. Past related tweets are also rectified by using this function. Just because of this method result set correctly identifies the event.

The concept of scoring function has been introduced which is calculated as:
ϕ(F)= α(T)β(C)
α(T)=Tweet score
β(C)=check score
A naive-Bayes approach to this problem has been taken:
α(T) = {P(E|Tw)/P(Ē|Tw)} >1
This states that probability of tweet Tw belonging to event E. This is given by naive bayes.
Taking log on both sides

This can also be re-written as follows:

α(T) = { logP(E|Tw)-logP(Ē|Tw)>0 }

$$\alpha(T) = \begin{cases} 1 & \log P(E|Tw) - \log P(\bar{E}|Tw) > 0 \\ 0 & \text{Otherwise} \end{cases} \quad (1)$$

Candidate tweet is Tw and the event class E.

A tweet is considered when the probability of belonging to the event class (E) is greater than the probability that it does not belong to the event class ($\bar{E}$).

$$\beta(C) = \prod_{f \in F} \max_{k \in F^R} \delta_k(f) \quad (2)$$

In this the task is to assign independent scores to fields that allows to define $\delta_x(f)$ for any f and x ϵ {What, When, Where}.

Therefore, pattern-based classification are considered for the extraction of event `When' and `Where' fields.

Given F^R = {What, When, Where} are the set of important fields for an event, a field set F has all its required fields if and only if F^R is true.

No classification uses baseline approaches for all fields.

$\delta_{what}(f)$ =0.5

$\delta_{where}(f)$ =Check(f,Address)

$\delta_{when}(f)$ =check(f,Date/Time)

What classification uses baseline approaches except for what fields.

$\delta_{what}(f) = \hat{w}_{what} . x_f$

$\delta_{where}(f)$ =check(f,Address)

$\delta_{when}(f)$ =check(f,Date/Time)

What-When-Where classification uses multiplication baseline approaches for all fields.

$\delta_{what}(f) = \hat{w}_{what} . x_f$

$\delta_{where}(f)$ =check(f,Address)**.** $\hat{w}_{where} . x_f$

$\delta_{when}(f)$ =check(f,Date/Time)**.** $\hat{w}_{when} . x_f$

Function $\delta_x(f)$, which computes the score for the maximum belongings class for each field. The function check (f, re) where f is a field, and re is a set of regular expressions for field type x, returning 1 if a field f is considered a match for field type x, and 0 otherwise.

If $\{\phi(F)=\alpha(T)\beta(C)\}>1$ then, we consider the tweets otherwise throw away from the database.

Select query is used in step 3 is as follows:

Select * from tablename where tweet =”%Event category”.

e.g. select * from tb where tid=”%job”.

This query is used to select domain based events like job, marriage, admission, holiday, unknown etc. Domain based Event classification is obtained by using this select query. Tweets belongs to trash topics (unimportant) comes in unknown category if tweets doesn’t belongs to any event category.

VIII. BENEFITS OF THE PROPOSED CLUSTERING ALGORITHM

The Event Extraction algorithm leads to the following benefits:

a) *Human Screening is completely eliminated:* As all know that twitter data grows exponentially so, human screening consumes more time. Proposed algorithm provides the way categorization of events category with human screening.

b) *Past related data eliminated up to some extent:* There are chances that past related tweets are categorized by the earlier algorithm. Proposed algorithm provides the way by which past related data using joint scoring function is eliminated up to some extent.

c) *Less training is required to do this task:* More training is required for identification of past related tweets. Proposed algorithm requires less training to do this task. Scoring function is used for correct identification of events and check function is used a filter which passes only present and future based tweets only. Hence, training required for this task is less.

IX. KEY CHALLENGES

There are number of key challenges in extracting Major life events from tweets. Key challenges determine that the difficulties that arise or problems that are present in the system and these problems are solved in this work.

Challenge 1: Multiple Definitions for Life Events:

Event identification is a major problem. While many types of events (e.g., seminar presentation, House rent, Movie screening, Birthday party) are important for categorization of events, It is difficult to prepare a list of characteristics for major life events on which algorithms can rely for extraction or classification.

Challenge 2: Huge amount of Twitter Data:

The user-generated twitter data found in social media websites is extremely is massive and sound. The language used to describe events is different from people to people and ambiguous and social media users most probably discuss public news which is current and life events from their daily lives, for instance what they ate for dinner. Even for a predefined life event category, such as Engagement, it is still difficult to accurately identify events.

X. IMPLEMENTATION AND RESULTS

For implementation which is shown in Figure 5. Firstly, collect the data set from twitterdata warehouse and store the tweets in the database named as wampserver. Then, connect the database with netbeans8.0. Apply LDA to identify the domains by which tweets belongs and discard the tweets those are not required.

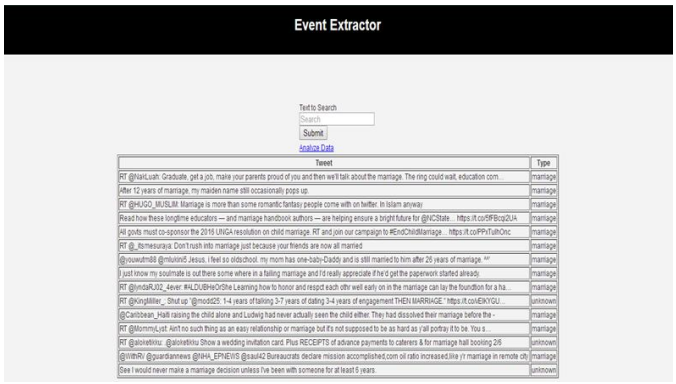

Figure 5. Screenshot of Event Extraction Implementation without Internet connection

In this implementation three domains are implemented that are marriage, job, admission and unknown rest for the others that do not belong to any of the three categories. So based on this approach we display the result with the help of Google graph.

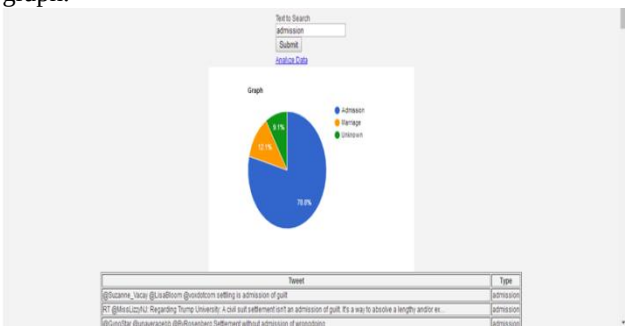

Figure 6. Screenshot of Event Extraction Implementation with Internet connection

Figure 6. show that above is the screenshot for the event category named as "admission". Results are analyzed by Google graph. This shows some tweets belonging to their respective event category in percentage.

There also needed of database to store all the tweets in which events are identified. Wampserver is used to store the database and for GUI interface. For Implementation, default browser is ucbrowser.

XI. COMPARISON BETWEEN EXISTING ALGORITHM WITH PROPOSED ALGORITHM

Here, in this main task is to show that proposed work is better than the earlier designed algorithm which is based on human screening and LDA. Algorithm proposed will improve the efficiency of results. The numbers of tweets which are identified by earlier designed algorithm in some of them are not correctly identified. Some past related tweets are not correctly identified by the earlier algorithm. More human screening is required to correct identification of such tweets. So, based on this proposed work correctly identifies such tweets. So, efficiency and accuracy is improved. The below figure shows the results are correctly identified than previous results. Figure 7. refers the graph which is designed according to the previous algorithm. Figure 8. refers the graph which is designed according to the proposed algorithm.

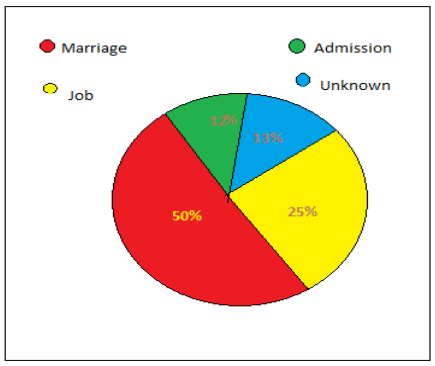

Figure 7. Graph of existing Algorithm

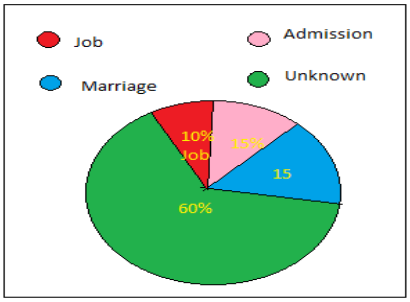

Figure 8. Graph of proposed Algorithm

XII. RESULT

The proposed algorithm that is given in this paper effectively discards the tweets that are not relevant for the users like past related tweets. This reduces the size of the dataset and makes the processing faster. There is no need of human interference in this algorithm; scoring will reduces the human screening requirement

XIII. CONCLUSION

In this section, Pipelined system based and an algorithm for major life event extraction from twitter is proposed. The main strategy preferred in this work is to fetch the more important, relevant category of Events from tweets. Because of particular interest in local major life events, this work is based on the identification and extraction of events on the internet. To achieve this goal, a couple of restrictions and scoring function, LDA topic modeling has been introduced. This extraction has the advantage that such technologies are adopted by many organizations like news agencies etc, the performance of event extractions will increase over time

without any extra labeling effort. This can be also adapted to crime investigation in various Field including Online Fraud Detection, Cell Phoning Crime investigation etc. The results are improved by using proposed algorithm described in Figure 3.

XIV. FUTURE WORK

Future work on Event extraction from twitter includes extension of this algorithm such that it also considers the semantic of the twitter data. A systematic method for the negative tweets with improved technique will be devised. It is expected that the proposed work will grow in the future into a complete system, which will give recommendations to users according to their interest and habits. The idea of proposed system is satisfactory and it is open to be implemented by future works. It is believed that the proposed algorithm will inspire the people who wish to develop applications on twitter data analysis and extraction.

XV. ACKNOWLEDGEMENT

A special thank to Dr. Parul gupta for her technical support to implement this model and also for useful comments, discussions, and suggestions regarding this approach. All data is extracted from twitterdata warehouse.

XVI. REFERENCES

[1]. Mohammad AL-smadi and Omar Qawasmeh "Knowledge-based approach for Event extraction from Arabic tweets",IJACSA,vol.7,No. 6, 2016.
[2]. John foley, Michael Benderky and Vanja Josifovski, "Learning to extract local events from the web"SIGIR, ACM 978-1-4503-3621-5/15/08, 2015.
[3]. Jiwei Li, Alan Ritter, Claire Cardie and Eduard Hovy. "Major life event extraction from twitter based on congratulations/condolences speech Acts",ICWSM, 11:438–4412015.
[4]. G. Katsios, S. Vakulenko , A. Krithara and G. Paliouras, "open domain event extraction from twitter", ACM 978-1-4503-1462-6 /12/08,2012.
[5]. Feifan Liu, Jinying Chen, Abhyuday Jagannathha, Hong Yu., "learning from Biomedical Information Extraction: Methodology Review of Recent Advances",DOI: http://dx.doi.org/10.1101/034397, 2016.
[6]. Abdur Rahman M.A. Basher, Alexander S. Purdy and Inanc Birol. "Event Extraction from Biomedical Literature",,Journal of Bioinformatics and Computational Biology,Vol. 8, No. 1 ,131–146, 2015.
[7]. Frederik Hogenboom, Flavius Frasincar, Uzay Kaymak, Franciska de Jong and Emiel caron, " survey of Event Extraction methods from text for decision support system", Decision Support Systems 55(1)256-269,2016.
[8]. Dr. D Ramesh, Dr.S.Suresh kumar. "Event Extraction from Natural language Text",In IJESRT in 2016.
[9]. Kolikipogu Ramakrishna, Vanitha Guda, Dr.B.Padmaja Rani , Vinaya Ch, "novel model for timed event extraction and temporal reasoning in legal text documents" , In Proceedings of the Fifteenth Conference on Computational Natural Language Learning, pages 49-57,2011.
[10]. H. Aliane, W. Guendouzi, and A. Mokrani, "Annotating events,time and place expressions in arabic texts." in *RANLP*, 2013, pp. 25–31.
[11]. J. M. Pawlowski, M. Bick, R. Peinl, S. Thalmann, R. Maier,D.-W.-I. L. Hetmank, D.-W.-I. P. Kruse, M. Martensen, and H. Pirkkalainen, "Social knowledge environments," *Business & Information Systems Engineering*, vol. 6, no. 2, pp. 81–88, 2014.
[12]. J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer *et al.*, "Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia," *Semantic Web*, vol. 6, no. 2, pp. 167–195, 2015.
[13]. W. Thamviset and S. Wongthanavasu. Bottom-up region extractor for semi-structured web pages. In ICSEC'14, pages 284{289. IEEE, 2014}.
[14]. E. Kuzey, J. Vreeken, and G. Weikum. A fresh look on
[15]. knowledge bases: Distilling named events from news. In CIKM'14, pages 1689{1698. ACM, 2014}.
[16]. R. Manjula and A. Chilambuchelvan. Extracting templates from web pages. In Green Computing, Communication and Conservation of Energy (ICGCE), 2013 International Conference on, pages 788{791. IEEE, 2013}.
[17]. Hristo Tanev, Maud Ehrmann, Jakub Piskorski and Vanni Zavarella, "Enhancing Event Descriptions through Twitter Mining", Sixth International AAAI Conference on Weblogs and Social Media,pages.
[18]. Jakub Piskorski and Roman Yangarbe, "Information extraction: past ,future ,present", DOI 10.1007/978-3-642-28569-1__2,© Springer-Verlag Berlin Heidelberg, in 2013.

Author's Profile

Ms. Monika Gupta pursed B.tech from M.D.U, Rohtak in 2014 and pursed M.tech from Y.M.C.A University in 2017. I am currently working as Assistant Professor in Y.M.C.A. University Faridabad.

Dr. Parul Gupta pursed M.tech from Y.M.C.A. University, Rohtak and her P.hd from M.D.U., Rohtak. She is currently working as Assistant Professor in Y.M.C.A. University, Faridabad.

