Volume 8, No. 5, May-June 2017

ISSN No. 0976-5697

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Implementation of Event Extraction from Twitter using LDA

Monika Gupta
Department of Computer Engineering
YMCA University
Faridabad, Haryana

Parul Gupta
Department of Computer Engineering
YMCA University
Faridabad, Haryana

Shruti Sharma
Department of Computer Engineering
]YMCA University
Faridabad, Haryana

Abstract: It is a simple approach for extracting events from twitter (tweets).Examples of events are national conferences, competitive exams, jobs opening, meditation camps etc. Tweets are the up-to-date and comprehensive collection of information about the current topic. It identifies the specific events in unstructured or semi-structured tweets. It transforms unstructured information in a collection of tweets into a structured database or in the form of retrieval calendar entries that contains information about specific event, time and location. The main goal of this paper is to explore different approaches for extracting the events. This paper focuses on usefulness of accurately extracting Major events.

Keywords: LDA, Twitter, events

I. INTRODUCTION

With the fast growth of social media, interest is increasing in detecting popular events from tweets. Event extraction is a work of identifying events from tweets or database of tweets. Each and Every day, hundreds of Megabytes of current stories are being added into the news archives of the major news agencies, containing much important and interesting news. The application of events includes time line formation, text summarization, FAQs etc. Events are also defined as predicates (statements) that describe the circumstances in which something holds true. Events may be expressed by means of verbs, adjectives, predicative clauses, or prepositional phrases. Event extraction refers to the task of discovering and saving structured representations of major life events from tweets with related attributes and properties, which are often, categorized by complex, and nested argument structures involving multiple entities. It is one of the atomic operations in detection and involvement among entities and other information in tweets or documents. Many of previous research on event extraction have focused on textual level extraction such as News articles, text summarization and Blogs, whereas few examples can be found on event extraction from noisy text such as tweets. For instance, tweets are short and self-contained which make them lack of useful information such as contextual information. The target of this research is to develop tools that extract and efficiently conclude major life events, the so-called (breaking news), extracted from social media. This task is useful for the professional journalists, it helps them to utilize social media as an information source helps to get a handle on with the lot of information. Twitter texts exclusive is its word count limitation which causes extensive usage of acronyms and other abbreviations. Event extraction combines knowledge and experience from a number of domains; it includes computer science, linguistics, data mining, and artificial intelligence. In the meantime, social networking sites such as Facebook and Twitter have become an important integral source of such type of information. Most of the time status messages contain useful information; they are much disorganized motivating the need for automatic extraction, aggregation and categorization.

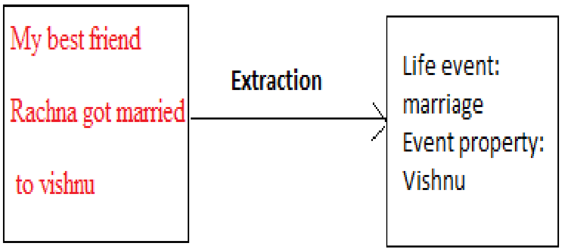


Fig.1

II. RELATED WORK

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]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]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] 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] 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] 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] 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).

Hristo Tanev, Maud Ehrmann, Jakub Piskorski and Vanni Zavarella[10]proposed the technique of Enhancing Event Descriptions through Twitter Mining through bigrams and unigrams with their frequency occurrences.

Jakub Piskorski and Roman Yangarber [11] proposed information extraction: past ,future ,present by using Named Entity Recognition(NER), Relation Extraction(RE) and Event Extraction(EE).

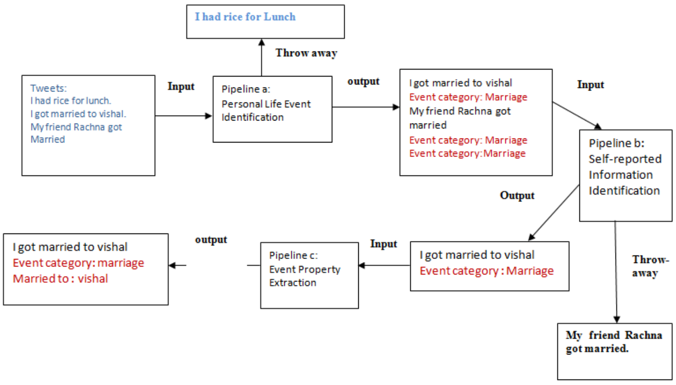


Fig 2

III. Pipeline Architecture of System

An overview of the components of the system is as follows: From Fig.2 Pipeline a first identifies the major life event category the input tweets speaks about and flush out the trash and unimportant tweets. Next, Pipeline b identifies whether the speaker is directly involved in the life event. Finally, Pipeline c extracts the event property. 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. Finally, pipeline c takes the output of second pipeline as input this pipeline extracts the events from tweets such as I got married to tom as event category marriage.

IV. APPROACH

An LDA based topic model is used to cluster the collection of tweets to find important categories of major life events in an unsupervised way. Then, associates each sentence with a single topic. Next some of authors manually identified the resulting major life event types inferred by the model, and manually assigned them labels such as ”seminar presentation”, “Movie screening”, ”admission” or ”marriage” and discarded irrelevant topics that uses a LDA-CLUSTERING+HUMAN-IDENTIFICATION strategy to identify public events from Twitter. LDA CLUSTERING + HUMAN IDENTIFICATION strategy to identify Major life events from Twitter In Fig. 3.
Similar strategies have been widely used in an unsupervised information extraction and selection preference modeling. In this also adopt a semi-supervised bootstrapping approach to identify reply seeds and event-related tweets.

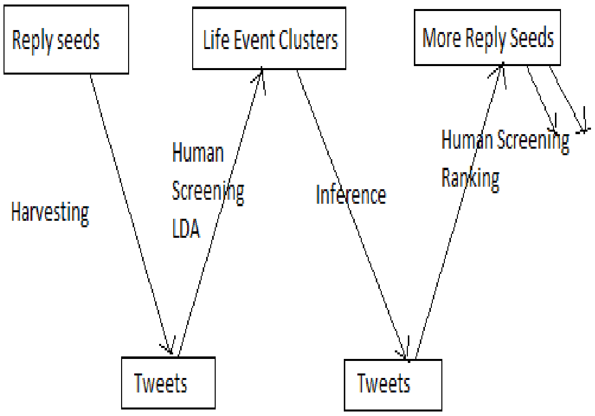


Fig 3

V. ALGORITHM

First of all, collect reply seeds tweets from twitter then apply a semi-supervised bootstrapping approach to expand the reply seeds and event-related tweets. Now, at this time apply the LDA modeling approach and by human screening identifies the relevant and trash topics and filter out the tweets based on this. Some major life events are manually identified on the topic modeling. Based on this we apply manually harvesting of tweets. Finally, we manually identify the best responses for the tweets. This approach requires more human identification and training. If training is not properly provided then there are more chances for the errors. Some tweets are more summarized in that case identification of such tweets required more identification and knowledge. Some tweets contains synonyms and some similar word to represent the same meaning so in that case more effort is required to find out such kind of tweets. This approach is not so good if the database of tweets is very large because there are some chances of miss out important tweets. This algorithm is simple because classification is done only with topic modeling other topics are treated as trash topics so we filter out such irrelevant topics. Here, this is the better way to extract the events from twitter.

The algorithm is as follows:

```
Input: Reply seed list E = {e}, Tweet conversation collection
T = {t}, Retrieved Tweets Collection D = ϕ.
Identified topic list L=ϕ
Begin
While not stopping:
    1.  For unprocessed conversation t ϵ T
        if t contains reply e ϵ E,
    •   add t to D: D = D + t.
    •   remove t from T: T = T - t
2. Run streaming LDA  on newly added tweets in D.
3. Manually Identify meaningful/trash topics, giving label to meaningful topics.
4. Add newly detected meaningful topic l to L.
5. For conversation t belonging to trash topics
    •   remove t from D: D = D - t
6. Harvest more tweets based on topic distribution.
7. Manually identify top 20 responses to tweets harvested from Step step6
    8. Add meaningful responses to E.
End
```

Fig . 4

VI. LIFE EVENT IDENTIFICATION

In this section, major life events are categorized into 42 different categories which are used in LDA approach. Some of them are listed in below table:

Categories are identified based on the words set in the table. There are some difficulties that is 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.

Table 1 Event Types Classification

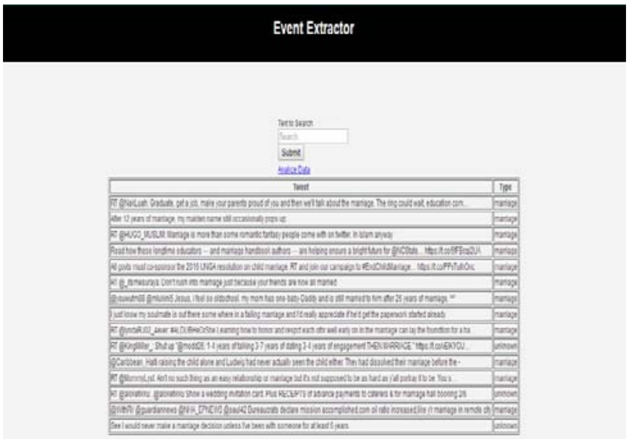
Human Label	***Tag Words***
Wedding & Engagement	Wedding, love, ring, engagement ,engaged, bride, marriage
Admission	Admitted, university, admission, college, offer, school
Exam	Passed, Exam, test, semester ,Exams
Research	Research ,presentation, journalism, paper, conference, go, writing
Movie	House,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

VII. IMPLEMENTATION AND RESULTS

For implementation, human screening is must required. 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+ human screening to identify the tweets that are relevant for us and discard the tweets those are not required. 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.

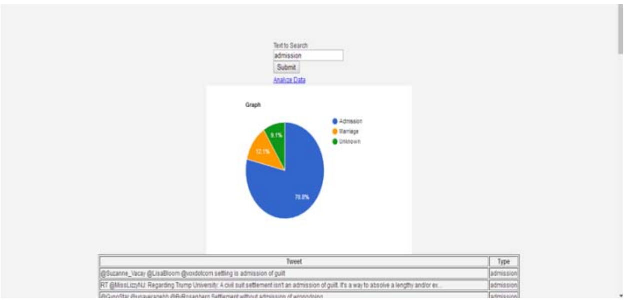
This shows the percentage of categories of major life event property. There are some difficulties that occur during implementation are more training is provided to implement this task.

This implementation requires more human screening so this requires more effort and time to identify each task. Cost will be increased as the data set increases and more training is required by this task is also increases. Here, some of the implementation screenshots are as follows:



Tweet	Type
RT @NaiLuah: Graduate, get a job, make your parents proud of you and then we'll talk about the marriage. The ring could wait, education com...	marriage
After 12 years of marriage, my maiden name still occasionally pops up.	marriage
RT @HUGO_MUSLIM: Marriage is more than some romantic fantasy people come with on twitter. In Islam anyway	marriage
Read how these longtime educators -- and marriage handbook authors -- are helping ensure a bright future for @NCState... https://t.co/9FBcqsZUA	marriage
All govts must co-sponsor the 2016 UNGA resolution on child marriage. RT and join our campaign to #EndChildMarriage... https://t.co/PPhTuflOhc	marriage
RT @_itsmesuraya: Don't rush into marriage just because your friends are now all married.	marriage
@ouwuthm88 @mlukin5 Jesus, I feel so oldschool. my mom has one-baby-Daddy and is still married to him after 25 years of marriage. ^^	marriage
I just know my soulmate is out there some where in a failing marriage and I'd really appreciate if he'd get the paperwork started already	marriage
RT @lyndaRJ00_&auer: #ALDUBHeOrShe Learning how to honor and respect each othr well early on in the marriage can lay the foundtion for a lu...	marriage
RT @KingMiller_: Shut up "@modd26: 1-4 years of talking 3-7 years of dating 2-4 years of engagement THEN MARRIAGE." https://t.co/AEKYOU...	unknown
@Caribbean_ Haiti raising the child alone and Ludwig had never actually seen the child either. They had dissolved their marriage before the -	marriage
RT @MommyLyd: Ain't no such thing as an easy relationship or marriage but it's not supposed to be as hard as y'all portray it to be. You s...	marriage
RT @alokationu: @alokationu Show a wedding invitation card. Plus RECEIPTS of advance payments to caterers & for marriage hall booking 2/6	unknown
@WilfriN @guardiannews @NIIA_CFNEWS @saul42 Bureaucrats declare mission accomplished.com oil ratio increased.like r\ marriage in remote cit	marriage
See I would never make a marriage decision unless I've been with someone for atleast 6 years.	unknown

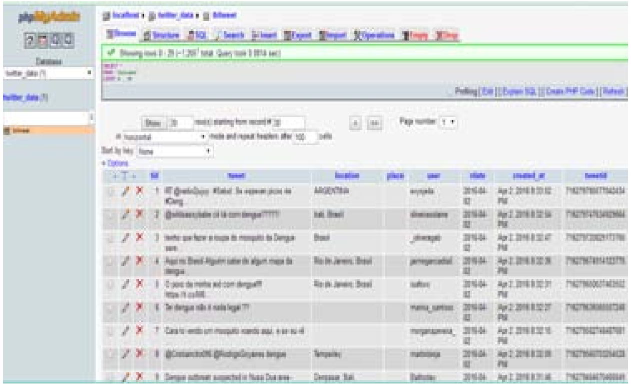
Fig.5



Tweet	Type
@Suzanne_Vacay @LisaBloom @voxdotcom settling is admission of guilt	admission
RT @MissLizzyNJ: Regarding Trump University: A civil suit settlement isn't an admission of guilt. It's a way to absolve a lengthy and/or ex...	admission
@GypoStar @unaveraneth @BjRosenberg Settlement without admission of wrongdoing	admission

Fig. 6

Figure 7 is the screenshot of database required to store the tweets. The database used to implement this algorithm is wampserver.



id	tweet	location	place	user	date	created_at	tweetid
1	[illegible]	ARGENTINA		[illegible]	2016-04-02	Apr 2, 2016 [illegible] PM	[illegible]
2	[illegible]	Itaí, Brasil		[illegible]	2016-04-02	Apr 2, 2016 [illegible] PM	[illegible]
3	[illegible]	Brasil		[illegible]	2016-04-02	Apr 2, 2016 [illegible] PM	[illegible]
4	[illegible]	Rio de Janeiro, Brasil		[illegible]	2016-04-02	Apr 2, 2016 [illegible] PM	[illegible]
5	[illegible]	Rio de Janeiro, Brasil		[illegible]	2016-04-02	Apr 2, 2016 [illegible] PM	[illegible]
6	[illegible]			[illegible]	2016-04-02	Apr 2, 2016 [illegible] PM	[illegible]
7	[illegible]			[illegible]	2016-04-02	Apr 2, 2016 [illegible] PM	[illegible]
8	[illegible]	Temperley		[illegible]	2016-04-02	Apr 2, 2016 [illegible] PM	[illegible]
9	Dengue outbreak suspected in Nusa Dua area	Denpasar, Bali		[illegible]	2016-04	Apr 2, 2016 [illegible]	[illegible]

Fig. 7

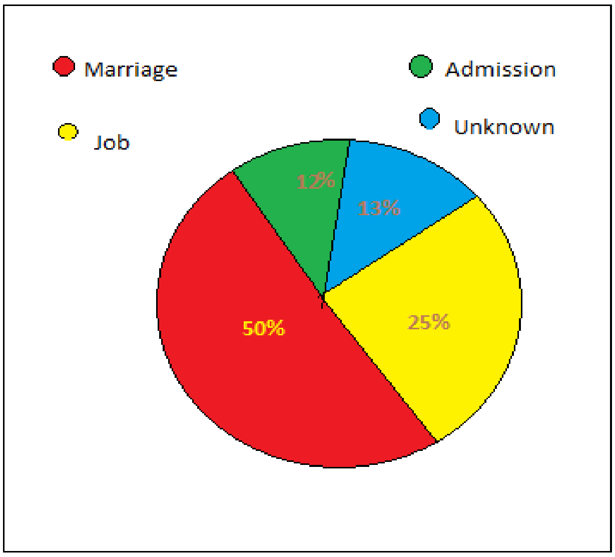


Fig. 8

VIII. KEY CHALLENGES

There are number of key challenges in extracting Major life events from tweets.

A. Challenge 1: Multiple Definitions for Life Events:

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

B. Challenge 2: Huge amount of Twitter Data:

The user-generated twitter data found in social media websites is extremely Big. The language used to describe life events is highly different from people to people and ambiguous and social media users frequently discuss public news 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 mentions.

C. Challenge 3: Insufficient of Training Data

Sufficient training data in this task for machine learning models is difficult.

IX. SHORTCOMINGS

There are several shortcomings of this approach.1. This approach is not suitable of very large data set. As we know millions of users are connected with social media so because of this data set is very huge. This approach can't handle large data set.2. Human identification approach is more error prone approach. By human screening there are more chances of Errors and time requirement is also high. 3. There are chances of missed out some important Data because of ambiguous definitions and lack of human training and knowledge. 4. More time consuming approach because of manually identify the major life events. Human identification takes more time required than machine based approach. 5.The system is only capable of discovering a few categories of life events with many others left unidentified because of ambiguous definitions and varied language based on people to people.

X. FUTURE SCOPE

This approach is not perfect for identify the major life events due to following ways:
1. Some important categories of major life events are left out because of ambiguous, noisiness, varied languages. 2. There are more chances of errors in each step of event extraction 3. Some people tweet in more comprehensive and short way because of this we cannot identify the major life event category. 4. More training is provided for correctly identification of events. So, because of the above reasons more work is required to correctly identification of events. Some other better technique is required to do this task. Tweets which are going in past tense are irrelevant for us so to filter out these kind of tweets require more training and human identification. This approach is not suitable for this type of extraction. Therefore, a more effective event extraction approach is required for to do this type of task.

XI. CONCLUSION

In this paper, Pipelined based system for major life event extraction from twitter is proposed. The Key strategy adopted in this work is to obtain the more important, relevant category of Events from tweets. Because of particular interest in local events, this work focuses on the identification and extraction of events on the open internet. To achieve this goal, we introduce a couple of restrictions and manual screenings, LDA topic modeling. This setup has the advantage that as more organizations adopt such technologies, the performance of event extractions will increase over time without any additional labeling effort. This paper aims to support manual identification and LDA approach for event extraction. This can be also adapted to crime investigation in various Fields including Online Fraud Detection, Cell Phoning Crime investigation etc.

XII. 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.

XIII. REFERENCES

[1] Mohammad AL-smadi and Omar Qawasmeh .2016. Knowledge-based approach for Event extraction from Arabic tweets ,IJACSA,vol.7,No. 6.

[2] John foley, Michael Benderky and Vanja Josifovski, 2015Learning to extract local events from the web.In.

[3] Jiwei Li, Alan Ritter, Claire Cardie and Eduard Hovy.Major 2015 life event extraction from twitter based on congratulations/condolences speech Acts.

[4] G. Katsios, S. Vakulenko , A. Krithara and G. Paliouras,2012, open domain event extraction from twitter.

[5] Feifan Liu, Jinying Chen, Abhyuday Jagannathha, Hong Yu. 2016, learning from Biomedical Information Extraction: Methodology Review of Recent Advances

[6] Abdur Rahman M.A. Basher, Alexander S. Purdy and Inanc Birol. 2015 Event Extraction from Biomedical Literature..

[7] Frederik Hogenboom, Flavius Frasincar, Uzay Kaymak, Franciska de Jong and Emiel caron, 2016, survey of Event Extraction methods from text for decision support system.

[8] Dr. D Ramesh, Dr.S.Suresh kumar. 2016 Event Extraction from Natural language Text,In IJESRT.

[9] Kolikipogu Ramakrishna, Vanitha Guda, Dr.B.Padmaja Rani , Vinaya Ch, 2011, novel model for timed event extraction and temporal reasoning in legal text documents.

[10] 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.

[11] Jakub Piskorski and Roman Yangarbe, 2013 information extraction: past ,future ,present, DOI 10.1007/978-3-642-28569-1__2, © Springer-Verlag Berlin Heidelberg,.