International Journal of Civil Engineering and Technology (IJCIET)
Volume 9, Issue 1, January 2018, pp. 320–329, Article ID: IJCIET_09_01_032
Available online at http://http://iaeme.com/Home/issue/IJCIET?Volume=9&Issue=1
ISSN Print: 0976-6308 and ISSN Online: 0976-6316

A SURVEY OF ENTITY IDENTIFICATION AND CLUSTERING USING TEXT PROCESSING FROM NEWSPAPER

Sridhar. Ranganathan
Associate Professor, School of Computing Science and Engineering,
VIT University, Chennai Campus, Chennai, India

Srivatsan. Kannan
Assistant Professor, School of Electronics Engineering
VIT University, Chennai Campus, Chennai, India

ABSTRACT

Newspaper industries are considered as the main path to face the audiences' expectations. It covers a huge range of applications in different fields such as business, politics, art and sport, which focuses on the following structures, opinion columns, reviews of local services, crosswords, obituaries, birth notices, weather forecasts, comic strips, editorial cartoons, and advice columns. Newspapers company is projected to contribute to the further development and improvement of Named entity identification systems with a focus on historical content. This paper explains the study of different methods based on the procedures followed in digital text processing, i.e., entity identification and clustering for detection and identification of name, location or a company etc. from the newspapers.

Key words: Newspapers, Named Entity Recognition, Clustering.

Cite this Article: Sridhar.Ranganathan and Srivatsan.Kannan, A Survey of Entity Identification and Clustering Using Text Processing from Newspaper. *International Journal of Civil Engineering and Technology,* 9(1), 2018, pp. 320-329.
http://iaeme.com/Home/issue/IJCIET?Volume=9&Issue=1

1. INTRODUCTION

The world has stepped to the epoch of big data. Processing the large volume of text in an high quality way becomes a significant issue. With the rapidly rising power of e-documents, several techniques have been adapted to process the massive information, such as: analysis and automatic detection of topics and, text summarization, and retrieval of information. Information extraction involves the method of finding and understanding of limited range, but relevant parts of the documents. Based on this, planned representation of the related information is produced.

In general, a technique of human-computer interaction is called Natural Language Processing. NLP is an application of computer science field that depicts how computers can

be used to identify and classify digital format text. Basically, the history of NLP generally starts in the 1950s, although work was started in previous time also. It has four methods: statistical, symbolic, hybrid and connections. The foremost task of NLP includes

- Information Extraction
- Speech segmentation
- Automatic summarization
- Machine translation
- Named entity recognition (NER), etc

A significant task is to determine the important themes of a manuscript; wherever matters can be referred as sentences, words, named entities and concepts. The Entity is defined as an independent presence of things. Each entity possesses its own nature, that is, various entities have their definite attributes which are easily differentiated from the other objects. The name of Entity often represents the species, that have the same nature as other nouns. People mention that name for every entity, which is also called as NE (Named Entity).

A Named entity is anything about a name. Named Entity recognition is a proper sequence of identification of the name and its classification. NER is an important part in the process of extraction of Information. Several applications of NER are determined in different subdivisions such as Machine Translation, Automatic Indexing of documents, Question Answering, Information Extraction, Cross-lingual retrieval of Information, Text Summarization etc.

2. CLASSIFICATION OF NER SYSTEM

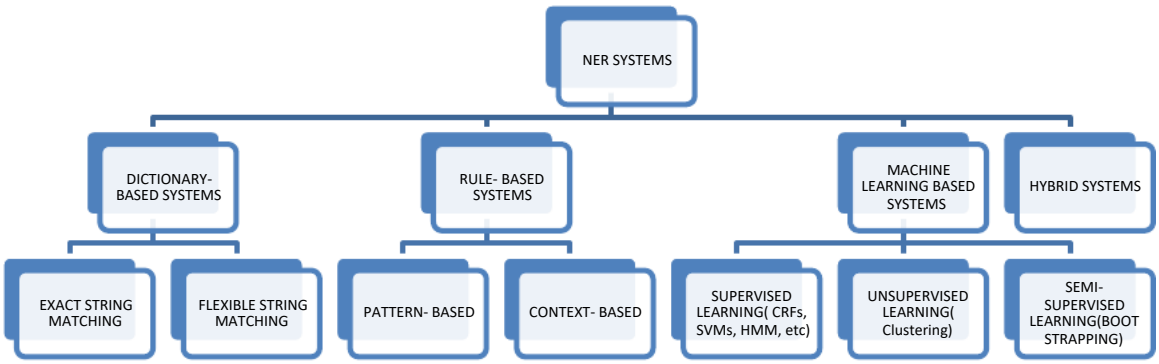

Figure 1 Structure of Named Entity Recognition System

The below mentioned structures are, Some of the classes of Named Entity found in NER

- Name of a Person
- Name of an Organization
- Time
- Location Name
- Abbreviation or Acronym
- Measure
- Term Name

The entities with the same categories have similar attributes, but they are different from the values of the property. Various types of entities have various properties. Any well-defined object is also an entity, which is different from various applications of information extraction. A user's interest can be also defined as an entity, such as people, products, etc., Moreover, objects that appear in the corpus can all be defined as an entity. Entities with various types have various attributes and information characteristics.

The description of the named entities is mainly considered with two structures, What and Who. Therefore, it is evident, that the conception of the named entity plays a key role in understanding of document and automatic extraction of information.

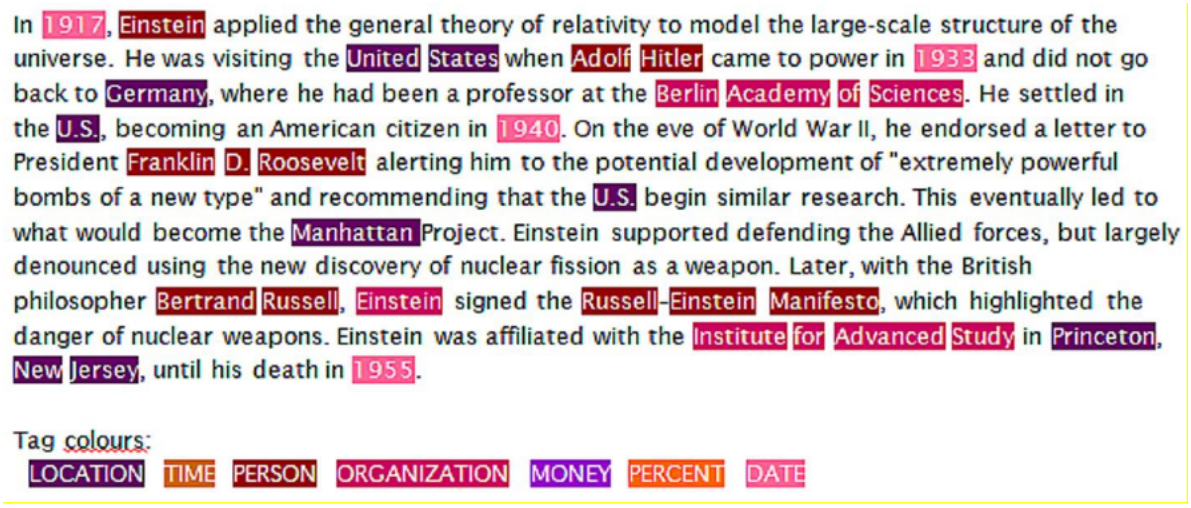

Figure 2 An example for Named Entity Recognition System output

An efficient and effective grouping method is required to extract consistent meaning from the newspaper. For this, clustering is used. Clustering, is a way of partitioning and grouping the data points of related properties [5]. It is not only a significant method for knowledge discovery and data mining but also a suitable method for information processing in several areas of application. Previous method of text clustering is based only on the occurrence of keywords (KW) in texts. Words contain those that denotes NE (named entities), which are called by the names such as locations, organizations, and people. In specific, news articles normally include the named entities, which are essential for the news contents.

3. RELATED WORKS

Sanjana Kamath. ET. AL [1] [2017] With the rise in easy use of data, the process of functional information extraction from available source has happened to the most significant activity across entire domains. When the data is provided as documents printed in natural language, the process of extraction of information becomes more complex. NER (Named Entity recognition) is a widely considered technique, for automatic extraction of theneeded information from amorphous natural language document collections. It is employed for both web applications as well as detached systems. NER is defined as the most significant steps in the NLP processfor text analysis. IT explains the following such as, fundamentals of NER, different algorithms adopted inthe process of NER and key applications and its issues in the applications of NER.

Sharnagat. Et. Al [2] [2014] explains the following techniques that were used for NERIt includes the following, supervised semi-supervised and unsupervised techniques. Named entity Recognition is a foremost step in NLP. This paper aims are to progress the system of

NER primarily for Indian languages. This study was considered as a numerical method of documentation. The detection of entities is carried out by discovering the names of entities such as, Name of a Person, Name of an Organization, Time, etc..The algorithms developed by considering the supervised learning processs were adopted to determine the Entities. Hidden Markov models, Decision trees, and SVM (support vector machines) are tinted as most widely and popularly used supervised learning algorithms for NER.The data sets in the Semi-supervised algorithms may be labeled or unlabeled. These algorithms generally begin with recognizing the small seed data and then move to huge amount of non-annotated data.

Pillai. Et. Al [3] [2013] explains how the languages contributes a major part in the process of NER, Language being the key goal of communication and assists in enabling machine type communication. This paper shows how language includes a vigorous role in talking, hearing, speaking, etc., The foremost part of NER is to classify and segment the various words in a text format into its succeeding categories like person name, place names, quantities, etc., It describes the 13 noun taggers for recognition of an entity like name of person, the names of location and organization names. The Hidden Markov model was adopted in supervised learning technique and also includes the statistical models with global learning method.

Kalyani Ramesh Pole. Et. Al [4] [2017] offers a diffuse linguistic topology space with a diffuse cluster algorithm to discover the basic related understatement and implication in documents related to web. The developed algorithm mines the functionality of network documents using methods of random conditional field and creates a widespread linguistic design space based on the relation of characteristics. The intrinsic links of words that possess the attribute to exist again in the hierarchy of chained semantic compound terms called as CONCEPTS, where a disperse measure of linguistic is given to each composite to estimate 1) the relation between the document features of a subject [10] and 2) the variation between the subjects. The data available on the Web can be combined into themes in the hierarchic structure [7], based on linguistic measures; users of the Internet can additionally discover the CONCEPTS of web content. Further the applicability of the algorithm in Web text fields, it includes various applications, such as bio informative, data mining, collaborative or content information filtering, etc.

Chien-Liang Liu. Et. Al [5] [2013] this work elaborates an algorithm known as fuzzy semi-k means which was based on concept of semi-supervised clustering. This algorithm is a derivative of K-means clustering model [6], and motivated by a Gaussian mixture model and an EM algorithm. Further, this algorithm offers the flexibility to use the various fuzzy membership functions to find the distance between data. This work makes use of the Gaussian weighting function to carry out the experiments, although the cosine similarity function is considered. This work was done with three data sets and comparative analysis was made. The output of the experiment reveals that fuzzy semi-k means was one of the best methods.

4. BASIC APPROACHES OF NER

Numerous approaches have been used in the Named Entity Recognition system are Handcrafted Approach/ Rule based, Automated or Statistical approach /Machine Learning, and Hybrid Model [1].

4.1. The Handcrafted or Rule based Approach

List Lookup Approach:

To classify the words, this NER system uses gazetteer. A suitable list has to be created by manually and it is fast, simple, and independent respective of the language. Their target process is easy to generate lists and it suits for lists in the a geographical index or dictionary.

The gazette should maintained. The uncertainty issue has not been overcome by the list lookup method.

Linguistic Approach:

The method of NER [8] uses few languages based policy and further experimental methods to categorize the words. The approach requires an expressive and rich system to produce the excellent output. It needs a superior knowledge of grammar and other language related convention. So the complete knowledge and highly developed skills related to the Language under consideration are needed to survive up with excellent quality rules and approaches.

4.2. Machine Learning Based Approach / Automated Approach

Hidden Markov Models (HMMs):

HMMs is called as a generative model. It schedules a joint probability of label sequence and paired observation. To expand the joint likelihood of training sets the parameters are qualified.

$$P(A, B) = \Pi i\ P(Ai, Bi)\ P(Bi, YBi\text{-}1)$$

It uses the Viterbi Algorithm, Estimation-Modification method, forward-backward algorithm, for modelling. Its major premise is easy to realize and neat. Therefore, its implementation and analysis is simple. Consecutively to define joint probability over label sequence and observation, HMM needs to list all feasible observation sequences. Therefore, it includes a variety of considerations about data like Markov assumption, i.e. recent label depends only on the earlier label. In addition, it is inconvenient to characterize long term dependencies and multiple overlapping features. Several parameters have to be evaluated is massive. Hence it requiresbulky data set for process of training.

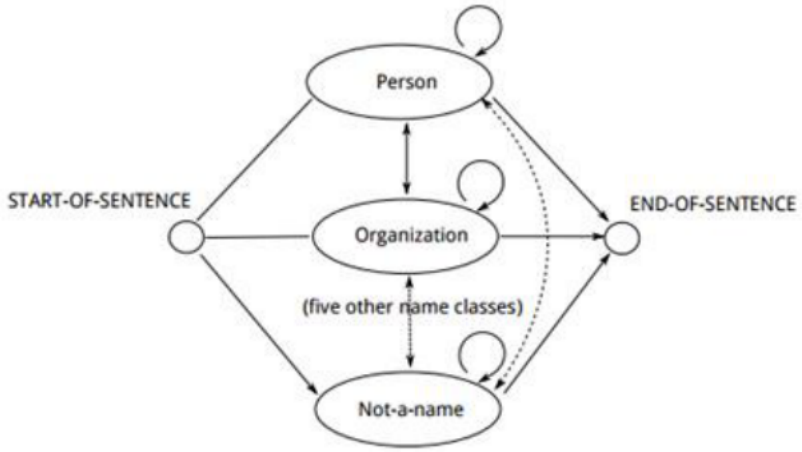

Figure 3 State Diagram for identifiers in HMM approach

Maximum Entropy Markov Models (MEMMs):

The MEMMs is a conditional probability sequence model. It can characterize various features of a word and can also manage long term dependency. This model is fully based on the maximum entropy principle, which explains that the slightest biased model where all the know facts should maximize entropy. Every source status has an exponential model that considers the observed feature as output and input a distribution over the likely next state. The labels of the Output relate to the states. It resolves the multiple feature representation problem and the issue of long term dependency faced by HMM. It generally enhances the accuracy, higher than HMM. It includes the issue of Label Bias. The probability transition leaving any

given condition must be added and the results should come to one. So, it is biased towards states with lower outgoing transitions. The circumstances with a single leaving state transition will not consider all the observations. To manage Label Bias Problem, the state-transition structure should be changed or a fully connected model has to start and let the training procedure decide an excellent formation.

Conditional Random Field (CRF):

This CRF falls under the type of discriminant probabilistic model. It has entire pros of MEMMs exclusion of the label bias issue. It is also referred as random field and they are undirected graphical models (also know as random field) which is used to estimate the conditional probability of values on the defined output nodes gives the values allocated to other defined input nodes.

CRF is of two types:

Higher order CRF: The basic CRF model can be moved to the higher order derivative by raising the number of the sequence of labels. In this model, labels are made dependent on fixed numbers of variables. If the number of variable becomes very large then training and inference become complex and hence alternative training and inference methods are applied.

Semi Markov CRF: This conditional random field method (semi-CRF), includes the dynamic-length segmentations (segmentation length may vary) of a label sequence. This offers the huge capacity of advanced order CRFs to build the wide dependencies of the sequence at a practical computational cost. The Application of CRF model:

- Named Entity Recognition
- Gene Finding
- Shallow parsing

4.3. Support Vector Machine (SVM)

SVM is the most popular supervised machine learning algorithms, used for binary classification in entire diverse data set and offers the best output, which includes the few data set with comprehensive algorithms which is used in multi-class issues. To work out a classification task by SVM, which was based on a supervised machine learning method, the task is usually computed with training data sets and testing data, which includes some data instances. Each instance in the training set contains one “target value” (class labels, where class label 1 for positive and class label -1 for negative target value and several “attributes” (features). The objective of a supervised SVM classifier technique is to generate a design which predicts the attribute’s target value. For each SVM, there are two data set, namely training and testing, where the SVM used the training set to make a model of classifier and classify testing data set relying on this model by considering their attributes or features.

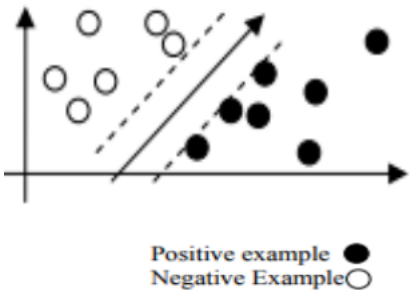

Figure 4 Linear Support Vector Machine Classification

Decision Tree (DT):

DT is a popular tool and powerful for prediction and classification. The main pros of the algorithm of Decision tree with respect to neural network is, it provides rules. The human can distinguish them or even openly use them in a database, contact language like SQL, therefore the records of the specified class may be a tree because the rules can be expressed during run time. DT is also a classifier, which includes the hierarchical tree model, where every node is considered as a leaf node-denotes the target attributes (class) of the expression value, or it can be a decision node that denotes some text to be considered as a single value of attribute with one branch and subtree for each probable outcome of the text. It is an inductive method to obtain knowledge on classification.

4.5. Hybrid Model Approach

In this Model, Machine Learning approaches and Rule Based approaches are used for getting more accurateness during the identification of NERs. Here, many combinations are considered.

- HMM approach and Handcrafted Approach/ Rule based
- CRF approach and Handcrafted Approach/ Rule based
- MEMM approach and Handcrafted Approach/ Rule based
- SVM approach and Handcrafted Approach/ Rule based

5. COMPARISON OF NER APPROACHES

For comparison, various methods have been chosen which includes the data set of MUC. These data collections of MUC were resulted from the articles of the air-accidents. The performance analysis of the NER includes calculation of three rates, namely Precision, Recall, and F. The scoring model was considered for the MUC and MET where P = No. of correct responses/ No. of responses, R= No. of correct responses/ No. correct in key, F=RP/1/2 (R+P).

Table 1 Hand made method results.

METHODS	R	P	F
List Lookup Approach	86	90	88.19
Linguistic Approach	85	87	86.37

From the above results, the system has reported comparatively better rate in all parameters.

Table 2 Machine Learning method results.

METHODS	R	P	F
HMM	89	96	92.20
MEMM	43.70	60.89	50.88
CRF	66.34	83.43	70.16
SVM	89.57	83.46	86.40
DT	89	92	90.44

From the above table, the variations in the results were caused by the amount of training datasets and different algorithms.

Table 3 Hybrid NER results.

METHODS	R	P	F
HMM & RULE BASED	N	N	94.50
MEMM & RULE BASED	92	95	93.39
CRF & RULE BASED	90	93	91.60
SVM & RULE BASED	85	93	88.80

From the above results, the system has reported high rates in all parameters. It is better to proceed with the hybrid models because it produces better results in terms of Recall, Precision and F.

6. CHALLENGES OF NER

Even though NER is considered to be a basic function of NLP [4], it includes some of the issues that are described below:

Ambiguity and Abbreviations - The key issue arises in discovering the named entities are language. Understanding the words which possess several meanings or words that can be a portion of various sentences. Another important challenge is separating the words that are alike from texts.

Spelling Variations-The vowels in English language, namely A, E, I, O, U includes a very significant character. It considers the text which does not create a wide difference in phonetics but make a huge variation in the mode of writing and its spelling.

Foreign Words – The Words which have not been used very often in recent days, or words that are not heard by most people, is another important limitation in this field. It includes the Words like names of a person, names of location, etc.

7. FACTORS RELATED TO NER:

Some of the basic factors that were related to NER are listed below:

7.1. Language Factor

Most of the research work has been done in English. Most of the domains in English have been explored. But this has confined the work to the particular language only. Language independence and multilingual are the crucial difficulties in this area. Languages like German, Spanish and Dutch have been studied in their own viewpoint. Japanese, Chinese, Swedish, French, Greek, Italian, have been researched in various literature. Survey on Hindi, Bulgarian, Korean, Danish, and Turkish are in progress. Even the Arabic started getting the awareness. But this works continues to be in its own boundary. Creating a multilingual NER is of main attention to the present situation.

7.2. Domain Factor

The two major aspects of a language have been ignored in the primary work of NER, firstly,text genre or text type like scientific, informal, technical, etc. Secondly, domain like sports, business, tourism, etc. Both the text type and the domain together structure the corpus to learn and examine the system. But, a solitary domain may have numerous corpuses within it. Hence, a particular language will have large corpora of text to evaluate. Relating the corpora of one domain to another is a foremost challenge.

8. APPLICATIONS OF NER

NER includes the wide range of applications in recent years. Some of them are listed below:

- It is successfully employed for automatic detection of events like crimes and disasters. The articles can be obtained using method of web based news aggregation,which is a software or an application which extracts the substances of the web).
- It is seen simultaneously in a several news articles, Synchronicity of names are considered, that is how frequent an Entity appears in a news article [2]. Nouns, proper nouns, etc. are also determined and used for populating databases.
- NER is widely used for identification of names of patient, address of patient, etc., from EMR (Electronic Medical Records). An automated tool is used to analyze the patient information. The information got from the EMR can be valuable for professionals in the medical field for added learning.
- It is used in social media domains for identifying the variety of Entities. Messages which are posted on Facebook and other social media may be unofficial in nature. Therefore, separating these named entities in social media is a complex assignment. The text differs from movie names, company names, brand names, etc.
- NER is acquiring a lot of attractiveness in analysis of social media as it includes the social networks such as Twitter and Facebook are used by people based on their opinions which form the root of several business processes (opinion mining).

9. CONCLUSIONS

This paper tells about the state of art Entity Recognition approaches [1], i.e. Rule based approach, Automated approach and Hybrid models. This study also reveals the challenges and applications of Entity recognition techniques. Natural language processing (NLP) and name entity recognition (NER) technique includes the clustering concepts is used to identify the latent semantic in documents. The development of hybrid models or approaches will give better results to gain the knowledge of the digital text or content. In a recent network world where a large amount of data is presented as documents in the natural language, NER has gained a major structure. It also shows, NER, a sub process of NLP, plays key role in the automated process of extraction of information.

REFERENCES

[1] Sanjana Kamath, RupaliWagh, “Named Entity Recognition Approaches and Challenges”, IJARCCE, Vol. 6, Issue 2, February 2017.

[2] Sharnagat, Rahul. "Named Entity Recognition: A Literature Survey." (2014).

[3] Pillai, Anitha S., and L. Sobha. "Named entity recognition for Indian languages: A survey." International Journal 3.11 (2013).

[4] Kalyani Ramesh Pole1, Vishakha R. Mote, “Name Entity Recognition and Natural Language Processing for Improvised Fuzzy clustering in Web Documents”, ICRISEHM, 2017.

[5] Chien-Liang Liu, Tao-Hsing Chang, Hsuan-Hsun Li, “Clustering documents with labeled and unlabeled documents using fuzzy semi-Kmeans”, Elsevier B.V.Fuzzy Sets and Systems 221 (2013) 48–64, 2013.

[6] I-Jen Chiang, Charles Chih-Ho Liu, Yi-Hsin Tsai, and Ajit Kumar, “Discovering Latent Semantics in Web Documents Using Fuzzy Clustering”, IEEE Transactions On Fuzzy Systems, VOL. 00, NO. 0, 2015.

[7] Athman Bouguettay, Qi Yu, Xumin Liu, Xiangmin Zhou, Andy Song, “Efficient agglomerative hierarchical clustering”, Expert Systems with Applications 42 (2015) 2785–2797.

[8] Tkachenko, Maksim, and Andrey Simanovsky. "Named entity recognition: Exploring features." KONVENS. 2012.

[9] D. Renukadevi, S. Sumathi, “Term Based Similarity Measure For Text Classification And Clustering Using Fuzzy C-Means Algorithm”, International Journal of Science, Engineering and Technology Research (IJSETR), Volume 3, Issue 4, April 2014.

[10] Rizzo, Giuseppe, and Raphaël Troncy. "Nerd: evaluating named entity recognition tools in the web of data." (2011): 1-16.

[11] R. Elankavi, R. Kalaiprasath and R. Udayakumar, A Fast Clustering Algorithm for High-Dimensional Data. International Journal of Civil Engineering and Technology, 8(5), 2017, pp. 1220–1227.

[12] Neeti Arora, Dr. Mahesh Motwani, A Distance Based C lustering Algorithm, International Journal of Computer Engineering and Technology, Volume 5, Issue 5, May (2014), pp. 109-119.

[13] R Gangadhar Reddy, M. Srinivasa Reddy, P R Anisha, Kishor Kumar Reddy C, Identification of Earthquakes Using Wavelet Transform and Clustering Methodologies. International Journal of Civil Engineering and Technology, 8(8), 2017, pp. 666 – 676