JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Approaches to Text Analytics and Mining in Multilingual Natural Language Processing

[1] Arjumand Masood Khan, [2] Khan Rahat Afreen
[1] Department of Computer Science and Engineering, Government College of Engineering, Aurangabad, Maharashtra, India
[2] Dr. Rahat Khan, Associate Professor, G. H. Raisoni International Skill Tech University, Pune, rahat.khan@ghristu.edu.in

Abstract: Natural language processing (or Computational linguistic is becoming the state of art in today's world. It has evolved many years ago in past 1960's. The task of NLP is understanding the natural human utter- ances in terms of speech or text, taking as input and giving proper response or output. Text mining also called as Text Analytics uses Natural language processing to transform unstructured corpus into standard and normalised documents or databases for further analysis by applying Artificial intelligence techniques and Machine learning algorithms.

IndexTerms **– NLP, URDU, ARABIC, Translation, Computational Linguistic**

I. INTRODUCTION

Effective and serious developments took place of Natural lan- guage processing in mid of 1970 to 1990 as various semantic and grammar rules for defining language, language structure, lexical & syntactical constructs of different languages (multilingual) came into effect. In present years we are facing extraordinary growth of indeterminate volume of data [1]. In recent period Text Analytics or Text Mining comprises of task such as information retrieval, text statistics and Machine learning [2], which is used to extract and process voluminous amount of data.

Hence scientific field of study of Natural Language Processing & Text Analytics combined with Machine learning and Data mining has helped for the generation and assessment of vast sources of corpus data (text) [1].

Text Mining & Text Analysis are often used as similar terms. In broad sense Text mining combine idea of statistics, linguistics and machine learning to create different models that can learn from training data & can process and predict new results from the pre- vious information. On other hand Text Analytics uses information from text mining & create graphs, plots and data visualisation. In most cases both processes are combined for compelling results.

While using Natural language processing an ordinary person should understand the syntax (what the word say) and semantics

(what the word mean) [3]. This can be further given for computa- tional linguistics for Text Analysis or Text Mining techniques. Semantic representation of a word or sentence in a language is important to avoid ambiguity in word or sentences [4]. India is a diverse nation. Many people of India speak & write different lan- guages, hence in this multilingual environment Morphology is essential part in multilingual setting in Natural language process- ing also called as Morphological parsing [5].

II. REQUIREMENT AND OBJECTIVE

Need of Morphological processing in Natural language process- ing is to overcome the problem related to semantics and the gram- mar rules of a language which can give rise to ambiguity in a language[6]. This also leads to problem of construction of compil- ers for language processing or translation to give the meaningful output.

Morphological parsing is the process of determining a smallest unit meaning called as morpheme for e.g. 'mangoes' is made up of word 'mango 'and suffix's'. Morphological parsing further connects to process of Syntax Analysis (parsing) then Semantic Analysis then Pragmatic Analysis and final Target representation.

Therefore, the research work should deal with extracting differ- ent features for different languages by applying different Meta classifiers algorithms & techniques.

III. NATURAL LANGUAGE PROCESSING TECHNIQUES

Following are some techniques which worked hand in hand for Natural language processing tasks or applications.

A. TEXT MINING

Text mining refers to the process of surveying & extracting an enormous collection of data which is in semi structured or unstruc- tured form. This process can incorporate various methodologies for analysing text for different research questions or task, out of which one is Natural language processing [7].

This organised data created by text mining can be used can be incorporated into data warehouses, business applications and intelligence for further descriptive, prescriptive or predictive analysis.

Text mining is very advantageous in terms of scalability, real time analysis & consistent criteria (maintaining consistency).

Variety of classification & Clustering algorithms are available for text mining for e.g. maximum likelihood for processing the large texts or documents.

B. TEXT ANALYTICS

Analysis of corpus (text) or data coming from Text mining for further procedure or unlocking unstructured data to give its mean- ing for finding out the pattern or theme is called Text analytics (TA). Text analytics refers to the use and need of customer require- ments. These requirements can be creating different data models, visualisation by creating charts, plots or graphs. Text analysis in Natural language processing means to transform text data into numerical format for subsequent analysis so that a user can get the desired output or response according to specific tasks of Natu- ral language processing.

Text Analytics can refer to different processes as Text Normal- ization or Text Pre-processing (e.g. dividing sentence into words, removing stop words, white spaces & find out the real word or pat- tern, e.g. ‘‘walked ? walk”, ‘‘processing ? process”.

Other methods can be Sentiment Analysis, Length Analysis, Lan- guage detection, Language translation, Text classification, Topic Modelling etc. Recently researchers are working more on Topic modelling, Sentiment analysis/opinion mining, Document classifi- cation and Document Clustering [6]. Nowadays people are more adhere to online chats or applications such as Facebook Twitter emails, tweets, blogs, Instagram. Hence Text analytics (TA) is to analyse your twitter account, millions & billions of emails, cus- tomer reviews in business applications/environment [8]. In this current pandemic situation due to COVID-19 online transactions, online teaching–learning communication, messaging, surfing, ecommerce is carried out in large amount, hence huge data/text has being processed and handled. Therefore, people want transac- tions or communication in their own languages generally including English. Hence forth Natural Language Processing with Text Ana- lytics is becoming state of art nowadays.

Several Text Analytics tools are also available to develop

Ontologies (study about the nature existence of the objects), vocabulary for Pattern based, Domain based, Rule based in terms of numerical methods or measurements in research.

These tools are responsible for transformation of document or text, linguistic processing, pattern recognition, identify or tag specific document section semantic tools to analyse medical/clini- cal texts/datasets.

Above Table 1 show various practice area related to desired products based on algorithms, methods or techniques of Text Analytics. The list can be more expanded or extended.

Table 1

Practice area of Desired product in text mining/text analytics.

Sr.No	Desired product	Practice area
1	Linguistic structure	Natural language processing
2	Marked sentences	Natural language processing
3	Topic assignment	Document classification
4	Document that match words	Information retrieval
5	Tagging text with predefined category	Natural language processing
6	Determining whether text is positive or negative	Natural language processing
7	Extracting main topics from corpus	Topic Modelling

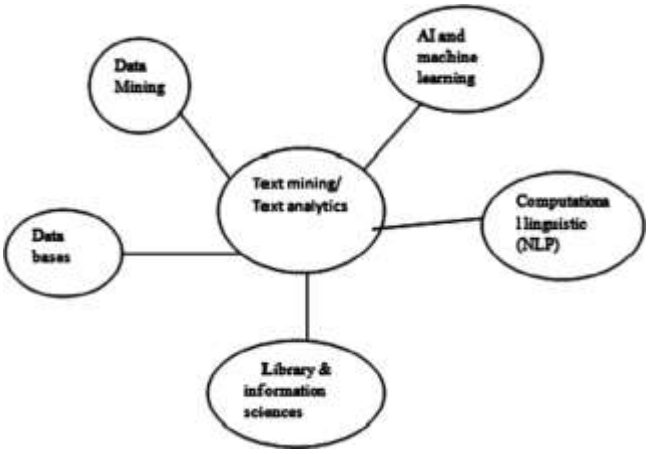

Fig. 1. Six fields related to Text mining / Text Analytics as Data mining, databases, Linguistics etc.

Fig. 1 shows six related fields related to different practice areas as shown in Table 1.

IV. MULTILINGUAL NATURAL LANGUAGE PROCESSING

Other than India other countries speak and write a specific lan- guage for e.g. Japanese, Chinese, English, but India is diverse coun- try, where many people speak, write and communicate in different languages according to culture and religion. In India too English is common language. Previous researchers focussed on computa- tional approach for the development of linguistic structure and varieties of techniques for speech recognition, speech synthesis, machine translation etc. Enormous development of this research is Google assistant and translator, Apple Siri and Amazon Alexa.

The above developments were best in English language, but researchers are working hard for the natural language spoke by people for e.g. Hindi language spoken by most of the Indian people. Google assistant can now process our language in Hindi too, but more development is required for low level languages as Hindi, Urdu, Sanskrit, Arabic etc for Text Analytics too.

One of the most spoken languages in United nation is Arabic. It is one of the languages of Holy Quran. Spoken and written dialects of Arabic are different. These dialects are called as LA (Levantine Arabic) & MSA (Modern standard Arabic), which is followed by dif- ferent people of different regions or countries, hence ontology and morphological analysis of a word or sentence or corpus is difficult to synthesis. Despite of many corpora available for us it is required for constant building of new corpora & updated it [9]. As in case of Low-level languages as Hindi & Arabic updated corpora should be constructed for better Text Analytics results. Therefore, in multilin- gual environment Text analytics can be manipulated with good corpora, algorithms or techniques or tools available for e.g. (Eng- lish + Arabic) Natural language processing.

V. SOFTWARE REQUIREMENTS

Following are some software requirements, tools and tech- niques for Text Analytics / Text mining in Natural Language Processing. Datasets are also available in good amount for multilingual systems.

A. DATASETS

- https://lionbridge.ai/datasets/best-arabic-datasets-for- machine-learning/
- https://old.datahub.io/dataset/jrc-names-ec
- https://data.mendeley.com/datasets/
- https://data.world/datasets/arabic
- https://sourceforge.net/directory/os:windows/arabic+dataset+classification
- https://www.semanticscholar.org/paper/
- .https://www.kaggle.com/mloey1/ahcd1
- .https://www.amia.org/education/webinars/i2b2clinical-nlp- datasets

B. TOOLKITS

- Stanford's Core NLP Suite – It includes tools for, and grammar parsing, named entity recognition tokenization, part of speech tagging.
- NLTK- Python language supports NLTK toolkit for same functions.
- Apache Open NLP.
- TACIT (The Text Analysis, Crawling, and Interpretation Tool (TACIT) is the graphical UI approach for tagging big data and
- provides state of arts for Text Analytics.

Various open source tools other than above tools are also avail- able for Natural language processing. These tools are mainly avail- able in Python programming. Variety of libraries and packages are available in Python language for stemming, lemmatization, mor- phological analysis, and tokenisation. For e.g. PorterStemmer pack- age is available in Python for stemming of a word.

C. TEXT BLOB AND VADER TOOL

Text blob is used for sentiment analysis; build on top of the NLTK (Natural language Tool Kit). It assigns polarity to word & approximate whole word sentiment as an average.

Vader (Valence Aware Dictionary and Sentiment Reasoner) is a rule-based model which particularly work on data of social media as tweets.

These tools are recently used other than NLTK, which are good for the beginners.

VI. CONCLUSIONS AND FUTURE WORK

Natural language processing includes Natural language under- standing and Natural language Generation which is to be evaluated through some computational linguistics. Text Mining and Text Analytics are the harnessing techniques for Natural Language pro- cessing. If corpus or text is plenty and if we are working in Mul- tilingual environment these techniques using some tools and algorithms are beneficial for metamorphosing and analyzing data to give predictable output.

These outputs can further be used for different Natural lan- guage processing applications and task. The work can be scaled by building the corpora for low level languages, updating it and using different platforms, Machine learning algorithms for text processing. Optimization of techniques and algorithms can also be included. Other than text work can be extended for images and videos too.

References

[1] Ashwin Ittooa, Le Minh Nguyenb, Antal van den Boschc, ''Editorial: Special issue on natural language processing and text analytics in industry ''HEC Management School, University of Belgium School of Information Science Japan Advanced Institute of Science and Technology, Japan - Division of Data Science, Ton Duc Thang University, Ho Chi Minh City, Vietnam Centre for Language Studies, Faculty of Arts, Radboud University Nijmegen Netherland 2016 http://dx.doi.org/10.1016/j.compind.2016.01.001.

[2] Anne Kao, Steve Poteet, Text Mining and Natural Language Processing- Introduction for the Special issue, SIGKDD Exploration 7 (1) (2005) 1.

[3] G. Miner, D. Delen, J. Elder, A. Fast, T. Hill, R. Nisbet, Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications, Elsevier, The Seven Practice Areas of Text Analytics, January 2012.
[4] Julia Hirschberg, Christopher D. Manning, Advances in natural language processing, vol 349, 17 July 2015.
[5] Daniel M. Bikel, Imed Zitouni, Multilingual Natural Language Processing Applications, Edited by IBM Press Pearson Upper Saddle River, ibmpressbooks.com ISBN-13: 978-0-13-715144-8, May 2012
[6] Arjumand Masood Khan, Dr. Rahat Afreen, Dr. Meghana Nagori, Text analytics on different corpus data for Multilingual System, Int. J. Sci. Eng. Manag. (IJSEM) 3 (2) (February 2018) All Rights Reserved © 2018 IJSEM 113, Aurangabad, India.
[7] R. Feldman, J. Sanger, The Text Mining Handbook: Advanced Approaches in Analysing Unstructured Data, Cambridge University Press, 2007.
[8] A. Balahur, M. Turchi, Comparative Experiments Using Supervised Learning and Machine Translation for Multilingual Sentiment Analysis. Association for Computational Linguistics. In Computer Speech and Language, 28 (1) 56-752014.
[9] Adil Rajput, Natural Language Processing, Sentiment Analysis and Clinical Analytics, Assistant Professor, Information System Department, Effat University An Nazlah Al Yamaniyyah, Jeddah 22332, Jeddah, Saudi Arabia February 2019 DOI: 10.1016/B978-0-12-819043-2.00003-4.

FURTHER READING

- Samir Tartir, Ibrahim Abdul-Nabi, Semantic Sentiment Analysis in Arabic Social Media, Department of Computer Science, Philadelphia University, Amman, Jordan. J. King Saud University – Computer Inform. Sci. 29 (2017) 229–233.
- Harshali B. Patil, B.V. Pawar, Ajay S. Patil, A Comprehensive analysis of Stemmers available for Indic languages, Int. J. Nat. Language Computing (IJNLC) 5 (1) (February 2016), School of Computer Sciences, North Maharashtra University, Jalgaon, India.
- Joseph M. Hilbe, Gary Miner, Robert Nisbet, Handbook of Statistical Analysis & Data Mining Applications, Nisbet, Elder & Miner 2009-06-05].pdf 31 oct 2008 Elsevier.
- Benjamin Bengfort Rebecca Bilbro, Tony Ojeda, Applied Text Analysis with Python: Enabling Language-Aware Data Products with Machine Learning, O'Reilly Media; 1 edition (July 10, 2018), ISBN-13: 978-1491963043.
- Steven Bird, Ewan Klein, Edward Loper, Natural Language Processing with Python, Analyzing Text with the Natural Language Toolkit, June 2009:O'reilly publication ISBN: 978-0-596-51649-9
- https://data-flair.training/blogs/python-tutorials-home.
- https://opensource.com/article/19/3/natural-language-processing-tools.
- https://towardsdatascience.com/text-analysis-feature-engineering- with- nlp- 502d6ea9225d.
- https://data-flair.training/blogs/python-stemming/.