INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume3, Issue4)
Available online at www.ijariit.com

Review on Text Classification by NLP Approaches with Machine Learning and Data Mining Approaches

Gurvir Kaur
Punjab Technical University
gurvir.kaur1990@gmail.com

Parvinder Kaur
Punjab Technical University
parvinderkaurcse@sus.edu.in

Abstract: Software Engineering and syntactic which is uncertain with the associations among PCs and regular dialects. In supposition, normal dialect changing over must be the perfect skill of human-PC interface. Characteristic dialect thankful is some of the time portray to as Artificial Intelligence-whole issues since common dialect acknowledgment appears to draw in broad learning about the outside world and the best to control it. NLP has huge have basic components with the field of computational phonetics and is regularly viewed as a sub-field of computerized reasoning. In this paper review on the different techniques of text, classification is discussed.

Keywords: NLP, Text Classification, Machine Learning.

I. INTRODUCTION

Natural Language Processing (NLP) is described as a stream of computer science and grammatical which is apprehensive with the connections among computers and natural languages. In supposition, natural-language converting must be the neat technique of human-computer interface. Natural-language grateful is sometimes describe to as Artificial Intelligence-entire issues because natural-language recognition seems to engage extensive knowledge about the external world and the aptitude to manipulate it. NLP has significant have common features with the field of computational linguistics and is often considered a sub-field of artificial intelligence.

NLP Terminology:

Token: The process of breaking input text into a unit like words, punctuation, numbers or alpha numeric, is called tokenization. These units are identified as tokens.

Sentences: The sentence is defined as a structured order of tokens.

Tokenization: It is defined as the procedure of breaking a sentence into its essential tokens. Tokenization is relatively easier for languages such as English due to the presence of whitespace. However, the mission is complicated for Chinese and Arabic languages, since there are no unambiguous restrictions.

Corpus: It is described as a text that commonly contains a lot of sentences.

Part-of-speech (POS) Tag: A word can be classified into a set of lexical classes like Nouns, Verbs, Adjectives, and Articles, to name of few others. Part of Speech identification is defined as a mark that represents a class such a lexical class - NN (Noun), VB (Verb), JJ (Adjective), AT(Article).

Parse Tree: A parse tree is described as a tree that is introduced over a definite sentence that represents the syntactic pattern of the sentence as described by a prescribed grammar.

Machine Learning:

Machine learning is a subfield of computer science (CS) and artificial intelligence (AI) that contract with the manufacture and research of systems that can gain knowledge from data to a certain extent that follows only clearly programmed commands. Moreover CS and AI, it has strong ties to statistics and escalation, which distribute both techniques and theory to the field. Machine learning is employed in a range of computing tasks were designing and programming precise, rule-based algorithms are infeasible. Example applications contain spam filtering, optical character recognition (OCR), search engines and computer vision. Machine learning, data mining, and sample identification are sometimes conflated.

Machine learning tasks can be of various forms. In supervised learning, the computer is offered with example inputs and their preferred results, given by a "teacher", and the goal is to learn the all-purpose rule that maps inputs to outputs. Spam filtering is an example of supervised learning, in meticulous allocation, where the learning algorithm is presented with email (or other) messages labeled in advance as "spam" or "not spam", to generate a computer program that labels hidden messages as either spam or not.
In unsupervised learning, labels are not mentioned to the learning algorithm, leaving it on its own to groups of similar inputs (clustering), density assumes or projections of high-dimensional data that can be visualized successfully. Unsupervised learning can be a purpose in itself (discovering hidden patterns in data) or a means an end. Topic modeling is an example of unsupervised learning, where a program is given a list of human language documents and is tasked to find out which documents cover on same topics. [11]

Text Mining:
Text mining is also acknowledged as data mining; indicate to the procedure of driving high worth information from text. The standard of data mining is a progression of raw and unstructured information take out significant information from text. It usually contains the technique of forming the contribution text, deriving samples contained by the structured statistics, and in consequence appraisal and examines of the output.
In content mining practice, initializing with the assembling of documents, a tool remove the particular knowledge or document and pre-process it. When it comes to the next phase i.e. text examine phase, where sometimes method used is repeated until the relevant information is extracted. Text mining and data mining are similar except the data mining tools which organize the structured data from the databases only whereas text mining extract information from semi-structured and structured data sets such as HTML files, emails etc. therefore text mining is better option to handle or organize online data for companies.
Text Mining tasks consist of:
Text Clustering
Concept/Entity Extraction
Text Summarization
Text Classification

1. Text Clustering:-Text Clustering (or Document Clustering) is described as a procedure of cluster examines to textual credentials. It has a function in routine certificate association, topic removal, and quick information recovery. Text clustering is usually measured to be a centralized technique. Text clustering consists of web paper clustering for investigating users is an example of text clustering. The content clustering might be of two types, online and offline. Online is generally controlled by effectiveness issues when compared to offline function.

2 Concept/Entity Extraction: Concept/Entity Extraction or (NER) Named separately identification is a sub-challenge of structured data mining that looks for the situated and organizes components in text into pre-described classifies such as the names of persons, places, situations of times, organizations, etc.

3 Text Summarization: Text summarization (or Automatic summarization) is the procedure of sinking a document from the data to generate a synopsis that contains the main significant points of the real document. With the quick expansion of online text or data, text summarization performs a vital function in text mining. A search engine like Google is an example of text summarization technology.

Commonly, extraction and generalization are two types of document summarization. Extractive processes work by choosing dividing up of current words, phrases, or sentences in the unique text to form the summing up whereas generalization techniques build an interior semantic demonstration and after that use ordinary language production method to make an outline that is convenient to what a human power produce.

Text Classification: The foundation of text categorization goes back to early 60's, but it became a major subfield in the early 90's. In the late 90's machine learning approaches were efficiently used for categorization of text. In 1998 SVM technique was used for categorization of text. In 1999 Maximum Entropy models were applied. In 1999 EM algorithm was also proposed by McCallum to train a hybrid model. In 2000 Multi-label classification is investigated of multiple topics and AdaBoost was used to enhance the multi labels classification. In 2005 Maximum Entropy models were extended to a new version multi-labeled Max Entropy Models for text classification. Content-based document management tasks had gained a major role in the information system field in recent ten years (2006-2015) because increased in accessibility of documents.

With the fast development of online information efficient recovery of several exacting knowledge is complicated with no excellent indexing and all over the report of document content. To grip and categorize the large text collection, text categorization may be the solution in text mining. In addition, Text Classification also called categorization is defined as a challenge defining without labeled documents into pre-described classes automatically.

II. LITERATURE REVIEW

Vishwanath Bijalwanet.al [1]: KNN based learning perspective is used for text classification. Text classification is the challenge of generally appropriating unlabeled credentials into pre-described categories. If a document or text belongs to accurately one class or category, it is known as single-label categorization task and if the document or text belongs to more than one or more class or category, it is called multi-label categorization task. The author firstly classifies the credentials with the use of KNN on

the basis of machine learning perspective comparing with Naïve Bayes and Term-graph approach by returning greater appropriate documents. An author achieves KNN displays the utmost correctness with correlation to the Naive Baye's and Term-Graph.
The drawback of KNN classifier is the time difficulty which is large but provides an improved correctness as a comparison to others. This mixture proves a goo0064 outcome than the typical amalgamation. At last, researcher finished in sequence reclamation application with the use of Vector Space Model to provide the outcome of the inquiry allowed by the client by proving the appropriate text.

Guansong Pang et.al [2]:- suggested a comprehensive cluster centroid on the basis of categorizer (GCCC) to use KNN and Rocchio via a clustering method. A design is grouped with Rocchio and KNN to formulate a comprehensive cluster centriod based model likewise to build convinced the scalability and uses of the GCCs model. The investigational outcome shows that GCCC proves constant positive presentation than KNN and Rocchio classifier. The disadvantages of GCCC is it's more prolonged than KNN and Rocchio.

Aixin Sun et.al [3]: - The author suggests a very clear measurable and non-parametric perspective for short text categorization. In general brief content are greatly noisier, briefer and sparser therefore to develop brief text representation author proposed to neat a brief text categorization. This prospective mimic's human categorization procedure for short text like tweets and comments. They chose the envoy terms as query words. Subsequent to that it finds for labeled text those perfect equivalent to the query terms. The researcher has used four perspectives and is appraised to choose the query terms: TF, TF.IDF, TF.CLARITY and TF.IDF.CLARITY. The recommend prospective achieves correctness with the base-line Maximum Entropy classifier. Investigational outcomes prove that TF.CLARITY enforce efficiently when more than three words are used in a query while TF.IDF.CLARITY enforce well when a single term is used in a query. The development becomes too small when five or more words are used in a query.

Youngjoong Koet. Al [4]:- developed text categorization by using powerfully class in a sequence to a word weighting mechanism. The researcher suggested a new strategy for various class content classifications. The out comes the prospective strategy applied set of information for word weighting for text categorization and enforced always on the data sets and KNN and SVM classifiers.

Eric H. Huang et al [5]: They introduce a latest neural network architecture which 1) learns word embeddings that superior capture the semantics of words by un-absorbing both local and global document perspective, and 2) accounts for homonymy and polysemy by learning numerous embeddings per word. They present the latest statistics with human judgments on pairs of words in sentential context, and appraise their model on it, displaying that their model outperforms competitive baselines and other neural language models.

QianXu et.al [6]: Inherent algorithms are the subjects of several scientific works. For example, in an investigation of inherent algorithms that are constructed for clustering ensembles, the genotypes, suitability occupations, and inherent operations are offered and it achieves that using inherent algorithms in clustering ensemble develops the clustering correctness. In this work, the k-nearest neighbor techniques are applied for the categorization of spam messages, and for the willpower of subjects of spam messages, clusters will be applied to a multi-document summarization technique presented in papers.

Sarwat Nizamani et.al [7] the correlation of superior topics likes multi-goal and ensemble-based developmental clustering, and the extending clustering is explained. Every of the algorithms that survey explained with reference to the fixed or a variable number of clusters; cluster-oriented or non-oriented operators; context-sensitive or context-insensitive operators; guided or unguided operators; double, integer, or real encodings; and graph-based representations. Clustering of spam messages simply means a usual grouping of thematically close spam messages. This issue becomes complex essential to carry out this process in a real-time mode in case of information streams as E-mails. There are separate methodologies that use unusual match algorithms for electronic documents in case of a significant magnitude of signs. When classes are defined by clustering technique, there is require of their support as spam messages continuously change, and the collection of spam messages replenishes. The new algorithm for the definition of criterion function of spam messages clustering problem has been offered in this paper. The inherent algorithm is used to solve the clustering issues.

Wen Zhanget.al [8]: In this paper author describes two tasks of text representation i.e. term weighting and indexing. In this paper author also have done a comparison of three methods (TF-IDF, LSI, multi-word) for text representations. This paper includes two documents i.e. Chinese and English are used to appraise the text representations method for text classification. The prime purpose of this paper is to read the efficiency and usefulness of separate text representation techniques. For content representation, two main tasks are indexing which is mainly concerned with statistical and semantic quality and weighting which is mainly concerned with term frequency (TF) and inverse document frequency (IDF).
The main objective is to research the efficiency of separate evidence technique in passage categorization. The author conducted the analysis to check the presence of the document representation technique i.e. TF_IDF, LSI and multi-word for phrase presentation. An analysis outcome established that in content classification, LSI has executed and performed better than other techniques in both document assembling. Moreover, although obtaining English documents LSI proved excellent presentation. The solution has displayed that LSI has positive syntactic and numerical superiority and is distinct by asset which LSI cannot generate particular authority for indexing.

Marco Pennacchiotti and Ana-Maria Popescu [9]: This paper addresses the challenge of user categorization in social media, with an application to Twitter. They naturally infer the values of user characteristics such as political direction or ethnicity by leveraging noticeable information. They engage a machine learning strategy which relies on an inclusive set of features derived from such user information. They report encouraging experimental outcome on 3 tasks with separate characteristics: political affiliation recognition, ethnicity identification and detecting affinity for a particular business.

mistake. Dalal and Mukesh A. Zaveri [10]: This paper describes the general approach for automatic content categorization and researches current solutions to big issues such as dealing with undesigned text, managing huge number of aspects and selecting a machine learning method suitable to the text-classification application. Automatic Text Classification has significant applications in content management, contextual search, opinion mining, product review examines, spam filtering and text sentiment mining.

REVIEW TABLE

Author Name	Year	Technology Used	Description
Vishwanath Bijalwanet.al.	2014	KNNbased Machine Learning Approach	The author firstly classifies the credentials with the use of KNN on the basis of machine learning perspective comparing with Naïve Bayes and Term-graph approach by returning greater appropriate documents. An author achieves KNN displays the utmost correctness with correlation to the Naive Baye's and Term-Graph.
Guansong Pang et.al.	2013	Cluster Centroid based classifier	Suggested a comprehensive cluster centroid on the basis of categorizer (GCCC) to use KNN and Rocchio via a clustering method. A design is grouped with Rocchio and KNN to formulate a comprehensive cluster centriod based model likewise to build convinced the scalability and uses of the GCCs model.
Aixin Sun et.al.	2012	Short Classification	The author suggests a very clear measurable and non-parametric perspective for short text categorization. In general brief content are greatly nosier, briefer and sparser therefore to develop brief text representation author proposed to neat a brief text categorization. This prospective mimics human categorization procedure for short text like tweets and comments. They chose the envoy terms as query words. Subsequent to that it finds for labeled text those perfect equivalent to the query terms. The researcher has used four perspectives and is appraised to choose the query terms: TF, TF.IDF, TF.CLARITY and TF.IDF.CLARITY.
YoungjoongKoet. al	2012	Term Weighting Schemes	Developed text categorization by using powerfully class in a sequence to a word weighting mechanism. The researcher suggested a new strategy for various class content classifications. The out comes the prospective strategy applied set of information for word weighting for text categorization and enforced always on the data sets and KNN and SVM classifiers.
Eric H. Huang et al	2012	global context and multiple word prototypes	They introduce a latest neural network architecture which 1) learns word embeddings that superior capture the semantics of words by un-absorbing both local and global document perspective, and 2) accounts for homonymy and polysemy by learning numerous embeddings per word. They present the latest statistics with human judgments on pairs of words in sentential context, and appraise their model on it, displaying that their model outperforms competitive baselines and other neural language models.
QianXu et.al	**2012**	Non-Content Features	In this work, the k-nearest neighbor techniques are applied for the categorization of spam messages, and for the willpower of subjects of spam messages, clusters will be applied to a multi-document summarization technique presented in papers.
SarwatNizamani et.al	2012	Enhanced Feature Selection	The correlation of superior topics likes multi-goal and ensemble-based developmental clustering, and the extending clustering is explained. Every of the algorithms that survey explained with reference to the fixed or a variable number of clusters; cluster-oriented or non-oriented operators; context-sensitive or context-insensitive operators; guided or unguided operators; double, integer, or real encodings; and

			graph-based representations. Clustering of spam messages simply means a usual grouping of thematically close spam messages.
Wen Zhanget.al	2011	TF*IDF, LSI, and multi words	In this paper, the author describes two tasks of text representation i.e. term weighting and indexing. In this paper author also have done a comparison of three methods (TF-IDF, LSI, multi-word) for text representations. This paper includes two documents i.e. Chinese and English are used to appraise the text representations method for text classification. The prime purpose of this paper is to read the efficiency and usefulness of separate text representation techniques.
Marco Pennacchiotti and Ana-Maria Popescu	2011	Machine Learning Approach	This paper addresses the challenge of user categorization in social media, with an application to Twitter. They naturally infer the values of user characteristics such as political direction or ethnicity by leveraging noticeable information. They engage a machine learning strategy which relies on an inclusive set of features derived from such user information.
mistake. Dalal and Mukesh A. Zaveri	2011	Automatic Text Classification	This paper describes the general approach for automatic content categorization and researches current solutions to big issues such as dealing with undesigned text, managing huge number of aspects and selecting a machine learning method suitable to the text-classification application.

REFERENCES

[1] Vishwanath Bijalwan, Vinay Kumar, Pinki Kumari, Jordan Pascual "KNNbased Machine Learning Approach for Text and Document Mining", *International journal of database theory and application*, Vol.7, No.1,pp.61-70, 2014.

[2] Guansong Pang, Shengyi Jiang, "A Generalized Cluster Centroid based classifier for text categorization" *Expert Systems with Applications*, Vol.8, pp.2758-2765, 2013.

[3] Aixin Sun, "Short Classification using very few words", *International journal of computer* applications technology and research, Vol.2, pp.4503-1475, 2012.

[4] Youngjoong Ko, "A Study of Term Weighting Schemes Using Class Information for Text Classification", International journal of computer applications technology and *research*, Vol.2, pp.4503-1475, 2012.

[5] Huang, Eric H., et al. "Improving word representations via global context and multiple word prototypes." *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics, pp873-882, 2012.

[6]QianXu, Evan Wei Xiang and Qiang Yang, "SMS Spam Detection Using Non-Content Features", *IEEE intelligent systems*, Vol.27, No. 6, pp. 44-51, Nov.-Dec. 2012.

[7] Sarwat Nizamani, Nasrullah Memon, UffeKock Wiil, Panagiotis Karampelas, "Modeling Suspicious Email Detection using Enhanced Feature Selection", *Wall street journal*, Vol.14, April 2012.

[8] Wen Zhang, Taketoshi Yoshida, Xijin Tang "A comparative study of TF*IDF, LSI and multi words for text classification", *Elsevier in advanced engineering software*, Vol.38, pp.2758-2765, 2011.

[9] Pennacchiotti, Marco, and Ana-Maria Popescu. "A Machine Learning Approach to Twitter User Classification." *ICWSM* 11.1 (2011): 281-288.

[10] Mita K. Dalal and Mukesh A. Zaveri, Automatic Text Classification: A Technical Review, Volume 28– No.2, and August 2011

[11]Svetlana Kiritchenko, Stan Matwin, "Email Classification with Co-training", International journal of computer applications, Vol.56, No.10, pp.1-10, 2006.

