ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 6, June 2019

Analysis of different Machine Learning Models used for Text Classification

Swathi P

Lecturer, Dept. of Computer Science Engineering, Sreechaitanya Degree College, Sathvahana University,

Telangana, India

ABSTRACT: The automatic classification of text documents has become a significant examination issue nowadays. Legitimate classification of text reports requires information retrieval, machine learning, and Natural language processing (NLP) techniques. Our point is to zero in on significant ways to deal with automatic text classification dependent on machine learning techniques, directed, unaided, and semi administered. In this paper, we present an audit of different text classification approaches under the machine learning paradigm. We expect our exploration endeavors give useful bits of knowledge on the connections among different text classification techniques just as reveals insight into the future examination pattern in this domain.

KEYWORDS: Information Retrieval, NLP, Machine Learning, Automatic text classification

I. INTRODUCTION

The massive accessibility of online text records has given us a too massive measure of information. This accessible information must be sorted out methodically for its legitimate use. Precise association of information encourages simplicity of capacity, looking, and retrieval of pertinent text content for the low application. Text Classification is an effective strategy for sorting out text records into classes.The automatic classification of text is a significant issue in numerous areas[1]. It has numerous applications, for example, mechanized ordering of logical articles, spam separating, recognizable proof of report sort, creation attribution, robotized exposition evaluating, study coding, classification of news stories, and so on. This undertaking falls at the intersection of information retrieval, NLP, and machine learning. Information Retrieval is the finding of the reports which contain answers to the questions. Objective measures and techniques are utilized to accomplish the said goal. Natural Language processing is utilized to improve the comprehension of natural language by speaking to archives semantically. This assists with improving classification results[2].

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 6, June 2019

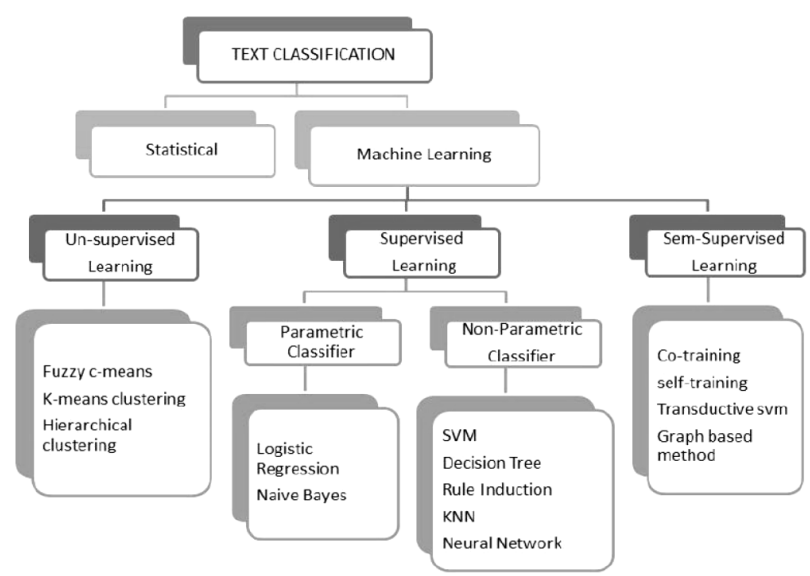

Fig 1: Representation of text Classifiers

Algorithms used to prepare text classification frameworks in information retrieval are regularly ad-hoc and ineffectively comprehended[3]. Precisely, next to no is thought about their speculation execution: their conduct on archives outside the preparation information. Machine learning techniques have a favorable position that they are better perceived from a hypothetical outlook, prompting execution certifications and direction in parameter settings.

II. SUPERVISED TEXT CLASSIFICATION ALGORITHMS

These algorithms utilize the training data, where each document is named by at least zero classifications, to become familiar with a classifier, which orders new texts[4]. A document is considered a positive model for all classes with which it is marked and as a negative example to all others.

K-Nearest Neighbor classifier

K-NN is a classification strategy where a given model is characterized by utilizing a dominant part vote of its neighbors. The perception is appointed to the class of its nearest k neighbors. The decision of K relies on the data given to the model [5]. For the most part, the more significant is the estimation of K; it diminishes the outcome of commotion or blunders in the classification. This calculation depends on the presumption that the characteristics of individuals from a similar class ought to be comparable. Accordingly, perceptions found near one another in covariate space are individuals from a similar class [6].

Naive Bayes Classifier

It is a probabilistic classifier ordinarily utilized in "ML." Via looking through the conditions among highlights, the presentation assumes an indispensable job. In this way, to improve the presentation, we can allot diverse weight esteems to highlights [7]. There are a couple of imperfections in highlight weighting techniques, for example bargained straightforwardness. So it continues to reduce the computational expense. The exhibition is subject to the exactness of the assessed probability terms. Precisely assessing the probability terms is hard for lacking training data [8].

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 6, June 2019

Decision Trees
It is a decision support device that utilizes a hierarchical structure of decisions. These decision trees plan to make a model that predicts the comparing classes by taking in fundamental decision rules from the training data[9]. These leaves represent the relating classification of the text documents, and branches speak to conjunctions of highlights that lead to these classifications.

Rule Induction
Another mainstream rule induction strategy is RIPPER. For the rules, the learning request is mandatory for viable arrangements since the arbitrary request of rules leads to blunders. Most usually utilized rule request improvement techniques are Particle swarm streamlining, Genetic calculation, and Simulated toughening. A couple of the enormous disadvantages of utilizing this method are rules being liable to late created guidelines, and rule learning happens progressively, further conceding the learning technique for another class[10].

Support Vector Machines
It is a statistical-based learning algorithm. This algorithm addresses the overall issue of learning to segregate among positive and negative individuals from a given class of n-dimensional vectors. It depends on the Structural Risk Minimization rule from the computational learning hypothesis. The SVM need both positive and negative training set, which are exceptional for other classification strategies. The SVM classification presentation stays unaltered regardless of whether documents that do not have a place with the support vectors are eliminated from the arrangement of training data

III. UNSUPERVISED TEXT CLASSIFICATION ALGORITHMS / TEXT CLUSTERING

In unsupervised clustering, we have an unlabeled assortment of documents. The point is to cluster the documents without additional information or mediation with the end goal that documents inside a cluster are more comparative than documents between clusters.

Hierarchical Clustering Techniques
In Hierarchical Clustering, the aim is to produce a hierarchical series of nested clusters. A diagram called **Dendrogram** (A Dendrogram is a tree-like diagram that statistics the sequences of merges or splits) graphically represents this hierarchy and is an inverted tree that describes the order in which factors are merged (bottom-up view), or cluster are break up (top-down view)[11].

Agglomerative Hierarchical Clustering
Agglomerative clustering is the most common type of hierarchical clustering used to group objects in clusters based on similarity. It has also known as *AGNES* (*Agglomerative Nesting*). The algorithm starts by treating each object as a singleton cluster. Next, pairs of clusters are successively merged until all clusters have been merged into one big cluster containing all objects.

1. **Single-link:** The single-link method defines two clusters *Ci* and *Cj* is similar to the two most similar documents. But
2. **Complete-link:** The complete-link method defines two clusters *Ci* and *Cj* is the similarity of the two least similar documents.
3. **Average-link:** The average-link method defines the similarity of two clusters *Ci* and *Cj* is the average of the pairwise similarities of the documents from each cluster:
   Text clustering aims to group text documents such that intra-group similarities are high

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 6, June 2019

Moreover, inter-group similarities are low. Document clustering has many application areas.Information Retrieval (IR), it has been used to improve precision and recall, and as an efficient method to find similar documents.

Polynomial regression

It is also a regression model representing the relationship between the observed data and the expected data as an nth degree polynomial. When the curve is built on a large number of observations, this method is more trustworthy.

K-means clustering

It intends to divide observations into K clusters. For example, in a supermarket, the items are clustered into different categories: cheese, butter, and milk, a group of dairy products.

Kohonen's Self Organizing Network

It uses a particular type of neural network called Kohonen's self-organizing network. The noveltyof the method automatically detects the number of classes present in the given set of text documents, and then it places each document in its appropriate class. The method initially uses Kohonen's self-organizing network to explore possible groups' locations in the feature space. Then it checks whether some of these groups can be merged based on a suitable threshold resulting in expected clustering. These clusters represent the various groups or classes of texts present in the set of given text documents. These groups are then labeled based on the frequency of the class titles found in each group's documents.

IV. CONCLUSION

It was verified from the study that among the unsupervised techniques, *k*-means andbisecting *k*-means perform the best in terms of time complexity and the quality of the clusters produced. On the other hand, support vector machines achieve the highest performance among the supervised techniques while naive Bayes performs the worst. Among the Semi-supervised techniques, the Graph-based algorithm performs well compared to co-training and Expectation-Maximization, but the hybrid algorithm of EM and naive Bayes gives better results.In many cases, hybrid algorithms outperform others and are gaining more research attention. However, there is a need to make more generalized text classification systems that will efficiently utilize massive amounts of unlabeled data. Such systems should also consider the nature of the input text documents.A few of the other vital issues are managing bulky classifications, data imbalance, document zones, feature selection, and performance-boosting.

REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in IEEE Conf. Comput. Vis. Pattern Recognit. IEEE, 2016, pp. 770–778. [Online]. Available: https://doi.org/10.1109/CVPR.2016.90
[2] Ankit Narendrakumar Soni (2018). Image Segmentation Using Simultaneous Localization and Mapping Algorithm. International Journal of Scientific Research and Engineering Development, 1(2), 279-282.
[3] Vishal Dineshkumar Soni. (2018). ROLE OF AI IN INDUSTRY IN EMERGENCY SERVICES. International Engineering Journal For Research & Development, 3(2), 6. https://doi.org/10.17605/OSF.IO/C67BM
[4] Ketulkumar Govindbhai Chaudhari. (2018). Analysis on Transport Layer Protocols and Its Developments. International Journal of Innovative Research in Computer and Communication Engineering, 6(9), 7707-7712. DOI: 10.15680/IJIRCCE.2018. 0609046.
[5] Ankit Narendrakumar Soni (2018). Data Center Monitoring using an Improved Faster Regional Convolutional Neural Network. International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, 7(4), 1849-1853. doi:10.15662/IJAREEIE.2018.0704058.
[6] Karunakar Pothuganti, Aredo Haile, Swathi Pothuganti,(2016)" A Comparative Study of Real Time Operating Systems for Embedded Systems" International Journal of Innovative Research in Computer and Communication Engineering, Vol. 4, Issue 6, June 2016.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 6, June 2019

[7] Ketulkumar Govindbhai Chaudhari. (2018). Comparitive Analysis of CNN Models To Diagnose Breast Cancer. International Journal of Innovative Research in Science, Engineering and Technology, 7(10), 8180-8187. DOI:10.15680/IJIRSET.2018.0710074.

[8] Vishal Dineshkumar Soni (2018). Prediction of Stock Market Values using Artificial Intelligence. International Journal of Advanced Research in Electrical Electronics and Instrumentation Engineering, 7(4), 1844-1848. doi:10.15662/IJAREEIE.2018.0704057.

[9] Ankit Narendrakumar Soni (2018). Application and Analysis of Transfer Learning-Survey. International Journal of Scientific Research and Engineering Development, 1(2), 272-278.

[10] Vishal Dineshkumar Soni .(2018). Internet of Things based Energy Efficient Home Automation System. International Journal of Innovative Research in Science Engineering and Technology, 7(3), 2924-2929. doi:10.15680/IJIRSET.2018.0703148.

[11] Ketulkumar Govindbhai Chaudhari. (2018). E-voting System using Proof of Voting (PoV) Consensus Algorithm using Block Chain Technology. International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, 7(11), 4051-4055. DOI:10.15662/IJAREEIE.2018.0711019.