IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol.-7, Special Issue-9, April 2019 E-ISSN: 2347-2693

A Review on Text Classification Algorithms and their Applications

G. Jasmine Beulah

Dept. of Computer Science, Kristu Jayanti College (Autonomous), Bangalore, India

Corresponding Author: jasmine@kristujayanti.com

DOI: https://doi.org/10.26438/ijcse/v7si9.2124 | Available online at: www.ijcseonline.org

***Abstract*—**Every day huge amount of text is generated online in vast quantities about the things happening in the world and in the minds of people. In order to generate meaningful business insights for organizations and analysts, the invaluable text data has to be mined. Extracting insights from unstructured text is costly and time consuming. Businesses use text classification to structure data in a fast and cost-efficient way to enhance decision making and automate processes. Text classification is a process of assigning tags according to the content with broad applications in sentiment analysis, spam detection, topic labeling and intent detection. This paper reviews the various text classification algorithms and their applications.

Keywords—Text classification, sentiment analysis, tags, spam detection, topic labeling, intent detection.

I. INTRODUCTION

Data is all around us at an increasingly explosive rate. Text data is a good example of unstructured information. Unstructured data is easily understood by human beings but it is difficult for machines to perceive. The enormous amount of text data is an invaluable source of knowledge and information (Wang & Chiang, 2011). As a result, there is a great need for effective algorithms to be designed and modeled so as to efficiently handle the enormous amount of data in a wide variety of applications.

Tremendous amount of text data is created everyday through social networks, news, emails, health care records, insurance records, etc and the area of Text Mining has gained importance due to the logarithmic increase in the volumes of data. The ever growing inflow of unstructured information is managed by organizations by text classification systems. The goal of text mining is to extract interesting or non-trivial patterns of information and use the knowledgeable information for strategic decision making (Dang & Ahmad, 2015). Within last few years, traditional data mining techniques and knowledge discovery have found a profound progress in the field of text mining. Text Classification helps the users to extract data from unstructured information and manages the operations like retrieval, extraction, summarization and categorization into supervised learning or unsupervised learning (Dang & Ahmad, 2015). Supervised Learning predicts the results within a continuous output and maps the input variables to some continuous function. In an unsupervised learning, the effect of variables on the data set is unknown and the output is unpredictable. Clustering algorithms are used to derive structure based on the relationships between variables in the data set. the applicable criteria that follow.

II. PROCESS OF TEXT CLASSIFICATION

The first and the foremost step in the process of text classification is identification of the data to be selected for training the machine learning algorithm. The unstructured data is then preprocessed using various preprocessing algorithms to reduce noise (Kotsiantis, 2007). The training data is further classified into supervised or unsupervised learning. The selection of algorithm to be considered and the nature and the type of the data set considered for the different machine learning algorithms is very crucial. Data mining techniques act as an experimental science as different parameters are tuned for different algorithms to choose the best classifier (Witten, Frank & Hall, 2011). The process of text classification is shown in Fig.1. Text classifiers or models can be used to organize structure and categorize data. For example, chat conversations can be organized by language, news articles can be organized by topics, support tickets can be organized by priority, social network can be organized by sentiments, etc.

Manual and automatic text classification systems can be done to interpret the content of text and categorize them accordingly. Human annotator interprets in manual text classification. In automatic text classification systems, machine learning, natural language processing and other techniques are used to automatically classify text in a faster manner and cost effectively. Automatic text classification system can be grouped into three different systems:

1. Rule based systems
2. Machine Learning based systems
3. Hybrid Systems

A. Rule based Systems

The Rule based systems uses linguistic rules to classify text and organize them. Each rule consists of patterns to semantically identify relevant categories of data based on their content. For example, news group can be classified into sports and politics group by defining two list of words that define and categorize each group. When a new line of text has to be classified, the number of sports related word count and politics related word count has to be calculated. The text is classified as sports related or vice versa based on the number of occurrences of the defined words. Rule based systems are human comprehensible but requires deep knowledge of the particular domain. Huge amount of analysis, time and testing are required for generating rules for complex systems. Rule based systems are not scalable as adding new rules can affect the pre-existing rules.

B. Machine Learning based systems

Classification of text is done by past observations by machine learning based systems. Pre-labeled examples are used as training data and the machine learning algorithm learns the different associations between the data and a particular output (i.e, tags) is expected for a particular input. Feature extraction is used to training a classifier with machine learning. Each text is transformed into a numerical representation in the form of vector. The machine learning algorithm is then fed with training data that contains feature sets and tags to produce the classification model as shown in Fig.2

Fig.1 Text classification process

With enough training samples, the machine learning model begins to make accurate predictions. Unseen texts are transformed into featured sets by the same feature extractor to get predictions on tags as shown in Fig.3.

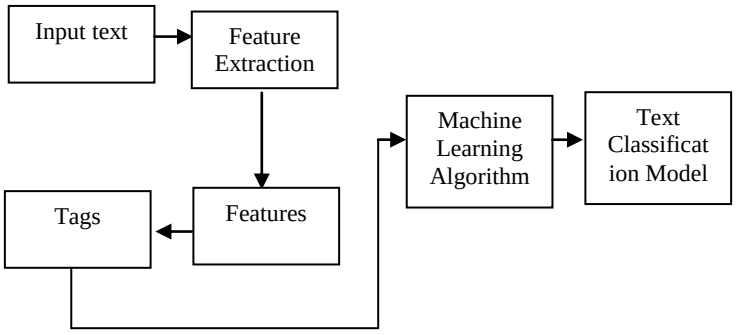

Fig. 2 Classification Model

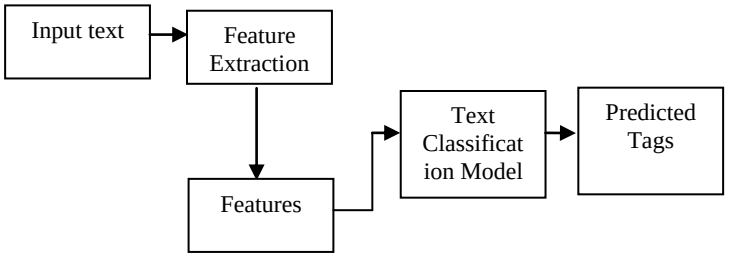

Fig.3 Prediction of Tags

Complex classification tasks can be solved using machine learning systems as they are easier to maintain and can always tag new examples to learn new tags.

III. TEXT CLASSIFICATION ALGORITHMS

The most popular machine learning algorithms for text classification are:

1. Naive Bayes
2. Nearest Neighbor Algorithm
2. Vector Support Machine
3. Deep Learning Algorithm
4. Artificial Neural Networks

A. Naive Bayes

Naive Bayes multiplies independent events assuming they are independent and it is based on Baye's theorem. (Witten et al., 2011). Any vector that means a text will have to contain information about the probabilities of occurrences of the word in the text within the texts of the given category so that the algorithm computes the likelihood of that text's belonging to a

particular category. Given a hypotheses (H) and evidence (E), the conditional probability is given by,

$$P(H/E) = P(E/H).P(H)/P(E)$$

B. K-Nearest Neighouring Algorithm(KNN)

KNN uses distance function to determine k members of the training set closest to the unknown test (Witten et al., 2011). KNN predicts the unknown test instance by the majority of k members. The classifier learns by comparing the given test sample with the similar training sets (Jiang, Pang, Wu& Kuang, 2012). (Jadhav & Channe, 2016)The KNN classifier follows the steps as:

Step 1: The value of K has to be initialized
Step 2: The distance between the input sample and the training samples has to be calculated.
Step 3: Sort the distance.
Step 4: The top K-nearest neighbors has to be considered
Step 5: Apply simple majority.
Step 6: Predict the class label with more neighbors for input sample.

C. Support Vector Machines(SVM)

Much training data is not required by SVM to achieve accurate results. SVM requires more computational resources to compute accurate results. Support vector machine algorithms use linear models to implement nonlinear class boundaries by transforming the instance space using a nonlinear mapping into a new space, a linear model constructed in the new space can then represent a nonlinear decision boundary in the original space (Witten et al., 2011). The principle of VC dimension from statistical learning and Structural Risk Minimization (SRM) (Singla, Chambayil, Khosla, & Santosh, 2011) is used in SVM. (Witten et al., 2011) SVMs are based on an algorithm that finds a special kind of linear model called the maximum-margin hyperplane. The instances that are closest to the maximum-margin hyperplane, the ones with the minimum distance are called support vectors as shown in the Fig.4 (Witten et al., 2011)

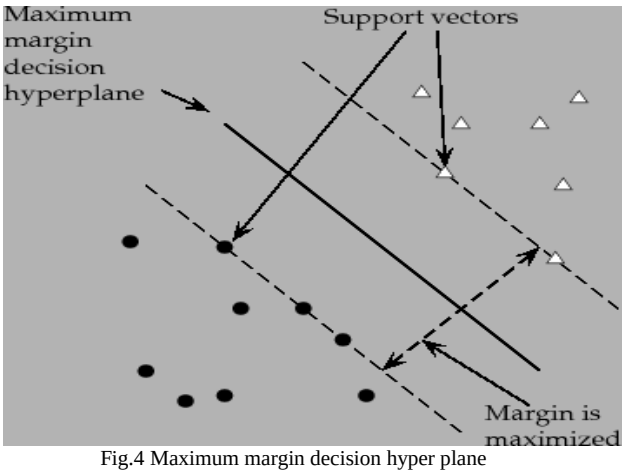

Fig.4 Maximum margin decision hyper plane

D. Artificial Neural Netorks (ANN)

Artificial Neural Networks is a computational model that has the ability to learn, memorize and generalize prompted research in algorithmic modeling of biological neural networks. Information processing system is the major element of ANN, where a large number of highly interconnected neurons work together to solve specific problems (Maind & Wankar, 2014). A simple artificial neural network is depicted in Fig.5 (Shahare & Giri, 2015).

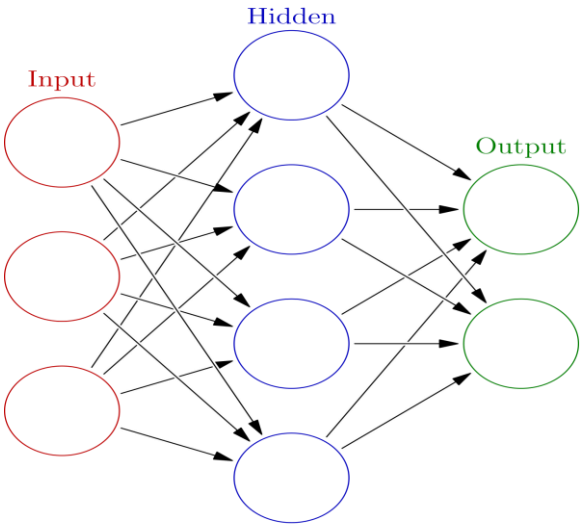

Fig.5 A simple artificial neural network

Through a learning process, ANN is configured for specific applications such as text classification or pattern recognition. The networks learn by examples and trained with known examples of the problem from which knowledge is to be acquired. The network can be effectively used to solve similar problems, when trained well.

E. Decision Trees

Tree based algorithms are considered to be the best algorithms and used in unsupervised learning. The tree builds classification or regression models in the form of a tree. They are adaptable at solving any kind of problem with high accuracy, stability, less data cleaning and ease of interpretation. Decision tree is mostly used for categorical or continuous input values. A decision tree is a tree which consists of nodes, links and leaf. Each node represents a feature or an attribute, each link represents a decision or a rule and each leaf represents a categorical or continuous outcome (Patel & Rana, 2014). The tree learns from a set of independent instances and is a probability distribution over all possible classifications. The dataset is divided into smaller subsets of data and a decision rule is incrementally developed.

IV. EVALUATION OF ALGORITHMIC METRICS

It is important to evaluate the performance of the different algorithms and determine the productivity of the data on which it was used to classify. (Yang, 1999) The classifiers can be evaluated by using binary classification, ranking performance, precision, F-measure, accuracy and error. Researchers can have a better idea on the different parameters to be considered while choosing the algorithm and also finding conditions for which the classifier works good. Over the last few decades, ensemble learning has also found to give classification accuracy over individual algorithms. Also, research on preprocessing techniques can help classifiers to predict better.

V. CONCLUSION

Machine learning algorithms on pattern recognition and text classification has found a profound development in extracting useful information for organizations. Review on the different popular algorithms used in text mining has shown the results to be quite accurate. Researchers have found great interest in processing and mining the huge amount of text generated on a daily basis.

REFERENCES

[1] Dang, S., & Ahmad, P.H ,"*A Review of Text Mining Techniques Associated with Various Application Areas",* International Journal of Science and Research (IJSR), Vol.4, No.2, pp.2461-2466,2015.

[2] Jiang, S., Pang, G., Wu, M., & Kuang, L," *An improved K-nearest-neighbor algorithm for text categorization. Expert Systems with Applications"*, Vol. 39, No. 1, pp. 1503–1509, 2012 https://doi.org/10.1016/j.eswa.2011.08.040271–350.

[3] Jadhav, S., & Channe, H,*"Comparative Study of K-NN, Naive Bayes and Decision Tree Classification Techniques",* International Journal of Science and Research (IJSR), Vol. 5, No. 1, pp. 1842–1845, 2016.

[4] Kotsiantis, S. B,*"Supervised machine learning: A review of classification techniques"*, Informatica, Vol. 31, pp. 249–268, 2007. https://doi.org/10.1115/1.1559160

[5] Patel, B. R., & Rana, K. K,*"A Survey on Decision Tree Algorithm For Classification*", International Journal of Engineering Development and Research, Vol. 2, No. 1, pp.1–5, 2014.

[6] Singla, R., Chambayil, B., Khosla, A., & Santosh, J, "*Comparison of SVM and ANN for classification of eye events in EEG",* Journal of Biomedical Science and Engineering, Vol.4 (January), pp. 62–69, 2011. https://doi.org/10.4236/jbise.2011.41008

[7] Shahare, P. D., & Giri, R. N,"*Comparative Analysis of Artificial Neural Network and Support Vector Machine Classification for Breast Cancer Detection*", International Research Journal of Engineering & Technology, pp. 2114–2119, 2015.

[8] Wang, T., & Chiang, H," *Data Mining: Practical machine learning tools and techniques",* San Francisco, CA, USA: Morgan Kaufmann Publishers Inc, 2011.

[9] Yang, Y. *"An Evaluation of Statistical Approaches to Text Categorization. Information Retrieval"*, Vol. 1, No. 1, pp. 69–90, 1999.https://doi.org/10.1023/A:1009982220290

