IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 17, Issue 1, Ver. IV (Jan – Feb. 2015), PP 80-84
www.iosrjournals.org

A Review: Text Classification on Social Media Data

Ms. Priyanka Patel[1], Ms. Khushali Mistry[2]
[1] *PG Student, Department of CSE, PIET, Vadodara, India, priyanka.23891@gmail.com*
[2] *Asst. Prof. Dept of CSE, PIET, Vadodara, India, khushali.mistry@gmail.com*

Abstract: *In today's world most of us depend on Social Media to communicate, express our feelings and share information with our friends. Social Media is the medium where now a day's people feel free to express their emotions. Social Media collects the data in structured and unstructured, formal and informal data as users do not care about the spellings and accurate grammatical construction of a sentence while communicating with each other using different social networking websites (Facebook, Twitter, LinkedIn and YouTube). Gathered data contains sentiments and opinion of users which will be processed using data mining techniques and analyzed for achieving the meaningful information from it. Using Social media data we can classify the type of users by analysis of their posted data on the social web sites. Machine learning algorithms are used for text classification which will extract meaningful data from these websites. Here, in this paper we will discuss the different types of classifiers and their advantages and disadvantages.*

Keywords: *Social Media Data, text classification, sentiment analysis, machine learning, classifiers*

I. Introduction

Social Media sites like Facebook, Twitter, LinkedIn and YouTube are the most popular sites among the Internet for all age group. These sites provide a social link with many people. Users of these sites are the one which shares, organize groups and provides useful information. When users post content on social media sites, they by and large post what they think and feel at that juncture. In this sense, the data gathered from online conversation may be more authentic and unfiltered than responses to formal research prompts. These conversations act as a zeitgeist for users' experiences [7]. All these Information contains a powerful meaning which classifies the type of users from their daily activities like daily posts, likes, comments, views, emotions with images, smiley and experiences. The social network provides a basis for maintaining social relationships, for finding users with similar interests, and for locating content and knowledge that has been contributed by other users. In social networks information filtering is used for avoiding the unwanted messages sharing or commenting on the user walls. Different types of machine learning methods are used for classification.

Opinion mining is a procedure to extract knowledge from the opinions that people share in web forums, blogs, discussion groups, and comment boxes [11]. In addition, opinion mining uses text mining and natural language processing techniques to make computer understand the expression of emotions. However, its main concern is to extract sentimental and emotional expressions from unstructured text [12]. Identifying the best method for classification is a critical task for sentiment analysis. Many of the approaches rely on database for sentiment analysis [13, 14].

Social media data provide great venues for students to share joy and struggle, vent emotion and stress, and seek social support. On various social media sites, users discuss and share their everyday encounters in an informal and casual manner. The development of social media sites among the people, it allows users to share their feelings and opinion. Our main aim is to review the different types of classifiers used for text classification and having an eye on their advantages and disadvantages.

In this paper Section 2 explains background, section 3 explains pre-processing in text mining, section 4 explains types of Classifiers, section 5 shows advantages and disadvantages of classifiers and section 6 conclusion.

II. Background

Web content mining is the procedure of extracting useful information from the web documents and which contains the generation of wrappers. Wrapper is a set of extraction rules to extract the data from the web pages, this can done either manually of automatically. The collection of data to be integrated may have different forms of content. This web content mining involves document tree extraction, data classification, and data clustering and finally labeling the attributes for results. Research activities are going on in information retrieval methods, natural language processing and computer vision [6].

Till now the Recommenders systems are used to suggest and improve the access to the relevant products like music, books and movies. A recommenders system by and large uses the content based filtering and collaborative filtering systems [1]. By applying the more than a few different text classifications methods used for

extracting the text from the social media sites. The system uniquely classifying the users interests by learning the information given in the profiles. Collaborative filtering technique works as filtering the information by collecting the user's preferences for particular item or opinion.

III. Pre-Processing In Text Mining

Gathered data from any social websites' can be in any one of the form (i) structured (ii) semi structured and (iii) unstructured. The data stored in databases is an example for structured datasets. The examples for semi structured and unstructured data sets include emails, full text documents and HTML files etc. Huge amount of data today are stored in text databases and not in structured databases. Text Mining is defined as the process of discovering hidden, useful and interesting pattern from unstructured text documents. Text Mining is also known as Intelligent Text Analysis or Knowledge Discovery in Text or Text Data Mining [15].

Gathered data from the social media website are just random in the structure and not even in well formed they just shared as the user feel at that particular moment. Now these gathered data is preprocessed by extracting proper and exact main terms. Text preprocessing steps include proper arrangement of documents. Preprocessing will increase the accuracy output, if done properly. There are two basic methods of text pre-processing: (a) feature extraction and (b) feature selection [3].

Text representation is the decisive task in the classification. It should be represented by collecting the set of features. Bag of words, document properties and contextual features are the types of features used. Text representation is underlying model of Vector Space Model (VSM). Bag of words are represented as the set of words presence in the documents and their allied frequency of weights [1]. Feature selection methods include the following:

- Document Frequency Threshold
- Information gain
- Mutual information
- Chi-square statistics

Feature selection is used to tumbling the high dimensional data space. Feature transformation methods include the embryonic semantic indexing. Selected features from the linear classifiers yields effective results.

IV. Types Of Classifiers

Classification is the separation or ordering of objects into classes [9]. There are two phases in classification algorithm: first, the algorithm tries to find a model for the class attribute as a function of other variables of the datasets. Next, it applies previously designed model on the new and unseen datasets for determining the related class of each record [10]. Text classification is to automatically assign the texts into the predefined categories. Text categorization mostly depends on the information retrieval technique such as indexing, inductive construction of classifiers and evaluation technique. In this machine learning, classifier learns how to classify the categories of documents based on the features extracted from the set of training data. Social content mining can be done on unstructured data such as text. Mining of unstructured data have hidden information and Text Mining is extraction of previously unknown information extracting information from different text sources. Social content mining requires application of data mining and text mining techniques [8].

Text is a kind of data in which the word attributes are sparse, and high dimensional with less frequencies on most of the words [8].To apply classification methods on text is difficult. The methods which are commonly used for text classification are follows:

A. Bayesian Classifier

The most commonly used classifier for Text classification. Basic idea behind this classifier is to find probability that to which class this document belong. Using this, we can understand the profiles by the feedback collected from various Social media sites. It is simple, but often outperforms more sophisticated classification methods. Maximum Likelihood estimates the parameters for the models. It requires small number of training to estimate the parameters. It Works well and efficiently in supervised learning. Here, the rank order of the pages will be rated. Text classification is based on calculating the posterior probability of the documents present in the different classes. Naïve bayes is based on Bayesian theorem with independence feature selection. Naïve Bayesian classification is used for anti spam filtering technique. It is divided in two different phases. The first phase has been functional for training set of data and the second phase employs the classification phase.

In Bayesian analysis, Prior Probability: It is a belief and based on previous experience. It is a ratio of number of single objects and number of total objects.
Likelihood: To classify a new object that this object belongs to which case.
Posterior Probability: The final classification is made by combining both sources of information i.e. Prior and Likelihood to form a Posterior Probability by Bayes rule.

Posterior Probability of X being a object α Prior Probability of total objects xLikelihood of X given objects.

B. Decision Tree

Decision tree is used for text classification it consist root node which contains all documents. Each internal node is subset of documents separated according to one attribute. Each arc is labeled with predicate which can be applied to attribute at parent. Each leaf node is labeled with a class. They designed a hierarchical decomposition of the data space. As per the attribute value it determines the predicate or a condition. In order to reduce the over fitting data, pruning is to be done. The listed splits are several different kinds of splits in the decision trees are available.

- Single attribute split
- Similarity-based multi-attribute split
- Dimensional- based multi-attribute split

They are implemented in the text context tend to be small variations compared to ID3, C4.5 for the purpose of the text classification [1].

C. K-nearest neighbor

K-NN classifier works on principle that is the points (documents) that are close in the space belong to the same class. It calculates similarity between test document and each neighbour. It is a case-based learning algorithm that is based on a distance or similarity function for pairs of observations, such as the Euclidean distance or Cosine similarity measures [2]. Many applications use this method because of its effectiveness, non-parametric and easy to implementation properties. However the classification time is long and difficult to find optimal value of k. The best choice of k depends upon the data. A good k can be selected by various heuristic techniques.

D. Support Vector Machine

Support Vector Machine finds out the linear separating hyper plane which maximizes the margin, i.e., the optimal separating hyper plane. Nonlinear separable case: Kernel function and Hilbert space. The SVM need both positive and negative training set as they are uncommon for other classification methods [3]. These positive and negative training set are needed for the SVM to inquire about for the decision surface that best separates the positive from the negative data in the n dimensional space, so called the hyper plane. The document representatives which are closest to the decision surface are called the support vector.

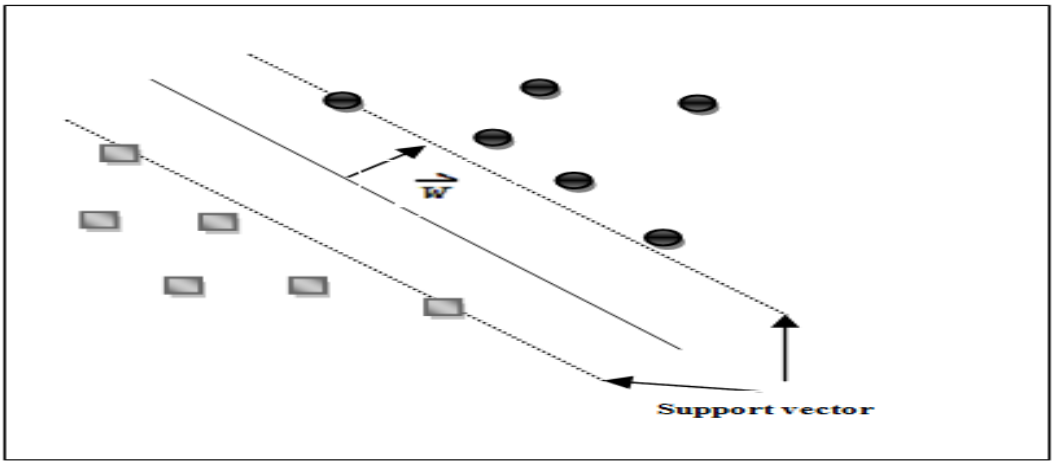

Fig. 1 Example of SVM hyper plane pattern [1]

The equation of the hyper plane for linearly separable space is WX+B=0
X is an arbitrary objects, W is a vector and B is constant learned from the set of linearly separable objects in the training documents. Vapnik proposed Classification algorithms for Support vector machines. Hyper planes are used to separate the two different classes of data.SVM can be operated on the pre classified documents [1].

E. Neural Network

The network comprises of a large number of highly interdependent processing elements (neurons) working together for solving any specific problem. Following is the Block diagram for neural network:

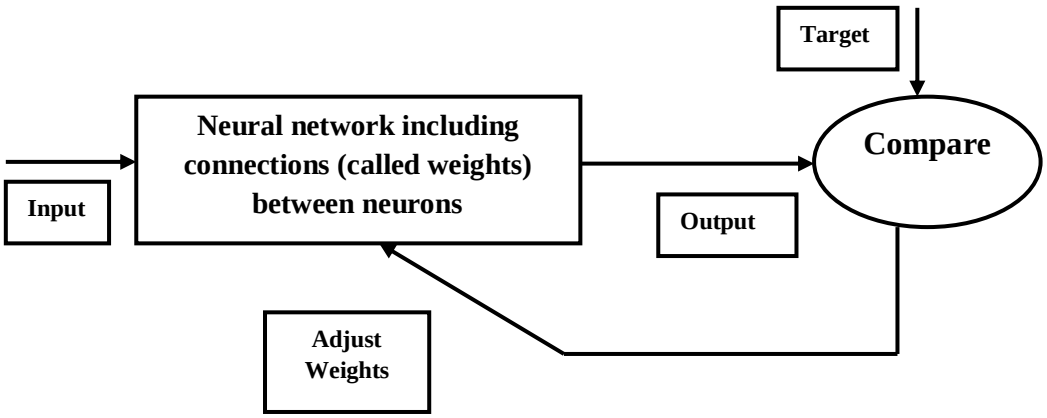

Fig. 2 Neural Network Block Diagram

As they have the ability to extract meaningful information from a huge set of data, neurons have been configured for specific application areas, such as pattern recognition, feature extraction, and noise reduction. In the neural network, connection between two neurons determines the authority of one neuron on another, while the weight on the connection determines the strength of the authority between the two neurons. There are two types of learning methods used in neural networks: (a) supervised learning and (b) unsupervised learning. In supervised learning, the neural network gets trained with the help of a set of inputs and required output patterns provided by a researcher [3].

The field of text mining is gaining popularity among researchers because of huge amount of text available via Social Websites in the form of blogs, comments, communities, digital libraries, and chat rooms. Neural network can be use for the logical management of text available on Social Websites.

F. Rocchio's

Rocchio's have to implement by using relevance feedback method. Synonymy means different have same or similar meaning. It can be addressed by manipulating the query or document using the relevance feedback method. In the relevance feedback method, here the user provides feedback which indicates relevant material about the specific domain area [3]. The user makes a simple query and the system in response with initial results in response to the query. Based on the result user decide is it relevant or irrelevant and then the algorithm may perform better. The relevance feedback method is an iterative process.

Ci= α * centroid ci − β * centroid ~ci [4] gives find similar method as of Rocchio is use in inductive learning process to find similarity between test example and category centroid using all feature .This algorithm is easy to implement, efficient in computation. The researchers have used a variation of Rocchio's algorithm in a machine learning context [5].

V. Advantages And Disadvantages Of Classifiers

Table 1 Advantages and Disadvantages of Classifiers [2][16]

CLASSIFIER	ADVANTAGES	DISADVANTAGES
Bayesian Classifier	• Work well on numeric and textual data. • Easy to implement. • Easy computation	• Conditional independence assumption is violated. • Performs very poorly.
Decision Tree	• Easy to understand. • Easy to generate rules. • Reduce problem complexity.	• Training time is relatively expensive. • One branch • Once a mistake is made at a higher level, any sub tree is wrong. • Does not handle continuous variable well. • May suffer from over fitting.
K-nearest neighbor	• Effective • Non-parametric • More local characteristics of document are considered comparing with Rocchio.	• Classification time is long. • Difficult to find optimal value of k.
Support Vector Machine	• capture the inherent characteristics of the data better.	• Parameter tuning • kernel selection

	• Global minima vs. local minima	
Neural Network	• Produce good results in complex domains • Suitable for both discrete and continuous data. • Testing is very fast	• Training is relatively slow • Learned results are difficult for users to interpret. • It may lead to over fitting.
Rocchio's	• Easy to implement • Very fast learner • Relevance feedback mechanism	• Low classification accuracy • Linear combination too simple • Various spelling correction techniques used.

VI. Conclusion

Electronic textual documents are highly obtained from the social websites. Large numbers of technologies are developed for the extraction of meaningful data from huge collections of textual data using different text mining techniques. However, Text pre-processing becomes more challenging when the textual information is not structured according to the grammatical convention. This review provides a thorough understanding of different text classifiers in the social networking websites.

From our review we concluded that different algorithms perform differently depending on data collections. . In this review we have seen the different classifiers and their advantages and disadvantages. Some algorithms do not perform well. None of them appears to be globally superior over the others.

References

[1]. K. Nirmala, S. Satheesh kumar and Dr. J. Vellingiri "A Survey on Text categorization in Online Social Networks" International Journal of Emerging Technology and Advanced Engineering Volume 3, Issue 9, September 2013.

[2]. Vandana Korde, C Namrata Mahender "TEXT CLASSIFICATION AND CLASSIFIERS: A SURVEY" International Journal of Artificial Intelligence & Applications (IJAIA), Vol.3, No.2, March 2012.

[3]. Rizwana Irfan, Christine K. King, Daniel Grages, Sam Ewen, Samee U. Khan, Sajjada. Madani, Joanna Kolodziej, Lizhe Wang, Dan Chen, Amma R Rayes, Nikolaos Tziritas, Cheng - Zhong Xu, Albert Y. Zomaya, Ahmed Saeed Alzahrani, And Hongxiang Li "A Survey on Text Mining in Social Networks, " The Knowledge Engineering Review, United Kingdom, (2004) pp.1-24.

[4]. Susan Dumais John Platt David Heckerman, "Inductive Learning Algorithms and Representations for Text Categorization", Published by ACM, 1998.

[5]. Michael Pazzani, Daniel Billsus "Learning and Revising User Profiles: The Identification of Interesting Web Sites", Machine Learning, pp. 313–331, 1997.

[6]. Ananthi.J "A Survey Web Content Mining Methods and Applications for Information Extraction from Online Shopping Sites", International Journal of Computer Science and Information Technologies, Vol. 5 (3), 2014, pp. 4091-4094.

[7]. Xin Chen, Mihaela Vorvoreanu, and Krishna Madhavan "Mining Social Media Data for Understanding Students' Learning Experiences," IEEE transactions on learning technologies, manuscript id 1, (2013), pp. 1-14.

[8]. Ms.S.Valarmathi, Mr.P.Purusothaman "A Survey on Web Content Mining Techniques and Tools", IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 1 Issue 6, August 2014.

[9]. G. K. Gupta, "Introduction to Data Mining with Case Studies." Prentice Hall of India, New Delhi, 2006.

[10]. P-N. Tan, M. Steinbach, V. Kumar, "Introduction to Data Mining." Addison Wesley Publishing, 2006.

[11]. Shahheidari, S.; Hai Dong; Bin Daud, M.N.R., "Twitter Sentiment Mining: A Multi Domain Analysis," Complex, Intelligent, and Software Intensive Systems (CISIS), pp.144-149, 3-5 July 2013.

[12]. Khan, K., B. Baharudin, and A.Khan. Mining opinion from text documents: A survey. In Digital Ecosystems and Technologies, 3rd IEEE International Conference on. 2009: IEEE: pp. 217–222.

[13]. Chaumartin, F., A knowledge-based system for headline sentiment tagging. In Proceedings of SemEval-2007, June 2007: pp. 422-425.

[14]. Valitutti, C.S.a.A., WordNet-Affect: an affective extension of WordNet. In Proceedings of 4th International Conference on Language Resources and Evaluation, 2004: pp. 1083–1086.

[15]. K.L.Sumathy,M.Chidambaram, "Text Mining: Concepts, Applications, Tools and Issues – An Overview", International Journal of Computer Applications (0975 – 8887), Volume 80 – No.4, October 2013.

[16]. Baijing, "Text Classification" http://www.iro.unmotreal.ca/~nie/ift6255/Classification.ppt