Review on Text Mining Algorithms

Shivani Sharma
M.Tech (CSE) Scholar
Department of Computer Science & Engineering
ABES Engineering College, Ghaziabad

Saurabh Kr. Srivastava
Sr. Assistant Professor
Department Of Information Technology
ABES Engineering College, Ghaziabad

ABSTRACT

Nowadays twitter microblog has become very popular in the conversation practice and in spreading awareness about various issues among the people. People share their short messages / tweets among their private / public social network. These messages are valuable for the number of tasks to identify hidden knowledge patterns from the discussions. Many research have been conducted on text classification. Text classification uses terms as features which can be grouped to vote for belongingness of a class. Text classification can be carried on twitter data and various machine learning algorithms can be used for feature based performance evaluation. In this context we have reviewed few papers taken from various sources like IEEE Xplore, ACM, Elsevier etc.

Keywords

Machine learning, Data mining, Features, Twitter.

1. INTRODUCTION

Data mining helps in gathering important knowledge from Data mining is the process in which valuable knowledge is extracted from large set of data. In data mining we discover useful and hidden knowledge from the database. Data mining is highly practiced in the companies nowadays to manage vast amount of textual data. These textual databases can be used for discovering the valuable knowledge related to products and services and for identifying users potentials and choices.

Information in data mining is broadly categorized into two main parts

a. Descriptive- Patterns are human interpretable.

b. Predictive- Interdependency of attributes.

Data Mining comprises of following steps

a. Data Cleaning- Data cleaning means removal of irrelevant data from the collection of vast amount of data.

b. Data integration- Data is collected from various different sources. Data integration means that data is combined in one common source from multiple sources.

c. Data selection- Data that is useful and relevant for the analysis is selected at this step.

d. Data transformation- Selected data is transformed into the form necessary for the mining operation.

e. Data Mining- It is the process of analysis of data to discover useful and hidden information. Data mining is the process in which valuable knowledge is extracted from large set of data.

f. Pattern evaluation- Patterns representing the knowledge are identified.

g. Knowledge Representation- It is the final step in which the knowledge that is discovered is represented to the user.

Data mining techniques come in two forms-

1. Supervised learning- Categories are known in supervised learning. Data is fully labeled data.

2. Unsupervised learning- Categories are not known. Learning process tries to find out categories. In unsupervised learning there are no predefined attributes. Hidden structure and relation among data is find out.

Classification is a major task to achieve through data mining algorithms. Classification uses machine learning to train the classifier. The goal of machine learning algorithms is to find patterns in the data. In textual context short text messaging (structured, semi-structured and unstructured) classification is the important research nowadays. In the textual context twitter research has become very popular. Twitter has become very famous microblogging tool and can be used to spread awareness or to spread messages among people in various areas. The areas can be healthcare, politics, education etc. Data collected from the twitter is very vast and unstructured and lot of useful information can be extracted from the twitter data. As twitter nowadays has become important source of communication among the people, they share their views on a particular topic through twitter. People can receive updates on important event through this microblogging tool. Thousands of tweets can be collected from twitter API and then data is classified into various classes for text classification.

1.2 Text Classification

Classification is a predictive task. It came from the area of machine learning. Classification is applied in many areas e.g. it can be used in classifying credit transactions as legitimate or fraudulent. It is a process in which we try to group records in the class. Large set of text data is collected and is assigned to one or more classes according to its subject type. After assigning the text data to the class we tokenize the data. Each word is called token of the class. So each class has attributes associated with it. These attributes are the features of the class. There are various tools that can be used for text analysis. Weka is one of them. Weka is a data mining tool that can be used for text classification. String data cannot be used in weka. So string data is converted into numerical or nominal form. The data can also be collected from twitter using twitter API and according to its subject type they can be categorized into various classes. Each class can have thousands of features associated with it. Many of the features are irrelevant i.e. they are useless. There are many different ways to reduce the features e.g. we can use stopwordlist to remove irrelevant features from the class. In text analysis terms are used as features of the class.

Any classification method consist of two data sets-

1) Training Data Set- This data set has a class that has labels in it. The class label is known. Model is created from the training data set which in future is used to predict the class of unknown records.

2) Testing Data Set- It is the data set which has similar description like training dataset but the values of attributes are not known. Values of these attributes are predicted by the previous observations of data. So we start using machine learning algorithms for finding the patterns in the data and then these attributes are associated with the value of the class. Testing data set is used to determine the accuracy of the model. Model is applied on testing data set.

If the data set is a big data set then we can have a better prediction model. We will never have model that is 100 % accurate. Some of the classification techniques are as follows-

a. Decision Tree Based Methods- Decision Tree is among the very popular machine learning techniques. It is liked by data miners because of user friendly results. It is simple, nonlinear and interpretable. Decision tree is based on top-down strategy. It is used to label unlabeled data. Decision tree is tree based structure. In decision tree we have root node, interior nodes, terminal nodes. Attribute is selected for root node and then branch is created for each possible attribute value.

b. Memory Based Reasoning- It is used to identify similar cases from experience. Application of memory based reasoning are fraud detection, medical treatments etc. In memory based reasoning neighbors are found such that it can be used for prediction and classification. Training data is a vast collection of records in which nearest neighbor to the unknown record can be used for prediction. The performance of Memory Based Reasoning totally depends on this training data set. Strength of memory based reasoning is that training dataset requires minimum efforts to get maintained. Results are easily understandable. But disadvantages of using this technique are that it is expensive and require large amount of storage.

c. Neural Networks- We all use computers everyday but sometimes computer fail us. So we like our computers to be smarter and user friendly. So we should try to make it more human. This would involve making computer think more like people. For it we should understand the working of brain. Working of computer is very simple. It take some inputs, process them and gets outputs. But the working of brain is not simple. Much research is going on in this field. Tiny components of brain are called neurons. There are billions of neurons in human brain. It is clear from the name that neural network consist of large number of processing units called neurons. Neurons are connected to each other with the help of link between them. Each link has weight associated with it. Signals are passed between neurons. Neural network can be used are for classifying patterns, grouping similar patterns etc.

d. Naive Bayes and Bayesian Belief Networks- Naïve Bayes is among the most effective algorithms. It is useful for large data sets. It is easy to build. I is easy to understand and debug. It is based on Bayes Theorem where knowing the value of an attribute does not tells anything about the value of other attribute. One of the advantages of Naïve Bayes is that a very small amount of training data is needed to estimate the parameters that are required for classification. Naïve Bayes is a conditional probability model.

e. Support Vector Machines- It is the most advanced and most popular machine learning technique. It is commonly used for text classification. It is the most successful method in machine learning. They are mostly used in high dimensional space. SVM is defined on both linearly separable data and nonlinearly separable data. The goal of SVM is to design a hyperplane. The hyperplane that has maximum margin from both the classes is best. Support vectors are the vectors that define hyperplane. It separates the two vectors into two linearly separable classes.

2. RELATED WORK

In the context of text classification AbdulkareemAlsudais et al. [1] used Random Forest to identify six location categories. They are active life, eating out, hotels, nightlife, shopping, and shows.16 tweet features are tested that belong to one of three groups: natural language processing (NLP) features, metadata features and establishments density (ED) features to predict category of tweet. He analyzed 43,149 reviews from Yelp**??** to train the classifier. He also examined two twitter data sets. First data set is the original data set that consist of 6359 tweets and the second data set consist of 2400 tweets which is uniformly distributed between six categories. These six categories are already discussed above. As there are two types of data set where one is training data set and other is testing data set. He has taken 60% tweets in training data set and 40% tweets in testing data set. He concluded that 74% of tweets were classified in the original data set and 77% of tweets are correctly classified in stratified data set. The goal in this paper was mainly to identify the type of location users are tweeting from.

Bo Pang Lillian Lee et al. [2] classified documents by overall sentiment. He does not classified documents on the basis of topics. He has taken movie review as data. Reviews were either positive or negative. Three machine learning methods were employed. Methods were Naive Bayes, Maximum entropy classification and Support vector machine. They do not perform well on sentiment classification as on traditional topic based categorization. Topic based categorization is to categorize documents according to its subject(e.g. sports verses politics). He concluded that sentiment categorization is more difficult than topic categorization. He examined the factors that made sentiment classification problem more challenging.

Evgeniy Gabrilovich et al. [3] described a class of text categorization problems that are characterized with many redundant features and developed a novel measure that captures feature redundancy, and use it to analyze a large collection of datasets and when no feature selection is performed then performance of text categorization with SVM peaks. They developed a measure that capture feature redundancy and then use this measure to analyze the vast collection of data.

Vandana Korde et al. [4] tried to introduce text classification, process of text classification as well as gave the overview of the classifiers. She compared some existing classifier on basis of criteria like time complexity, principal and performance.

Kamal Nigam et al. [5] showed that the improvement in the accuracy of learned text classifiers can be achieved by augmenting a small number of labeled training documents with a large pool of unlabeled documents. Based on the combination of Expectation-Maximization (EM) and naïve Bayes classifier an algorithm for learning from labeled and unlabeled documents is introduced. From the experimental results he found that classification error is reduced upto 30% by the use of unlabeled data. Combining both labeled and unlabeled training documents in EM performs well than taking labeled documents alone.

Thorsten Joachims [6] analyzed some properties of learning with text data. He identified why SVMs are appropriate for this task.He concluded from the experimental results that SVMs perform good on text categorization tasks. He found there is no need of feature selection in SVM. SVM do not require any tuning of the parameters.

Xiuju Fu et al. [7] analyzed the time series dengue data by using support vector machine classifiers and determined the time-lags and subset of climatic factors as effective factors by using the genetic algorithm influencing the spread of dengue. In the result they have shown that all the climatic factors can influence the Dengue process. Certain climatic factor is upper or lower bound is important. Meteorological data is provided by the National Environment Agency (NEA) of Singapore. Impact of climatic factors on dengue incidence trends was analyzed in Singapore. Temperature, Rainfall and Humidity are some key climatic factors that effects the magnitude of Dengue outbreaks. Some research works have been carried out to analyze the impact of Time Lags of climatic factors on Vector borne infectious diseases. In this paper there are few research studies in which data mining techniques have been applied in the analysis of dengue case data. Data mining analysis model is combined with SVM and Genetic Algorithm for investigating the relationship between climatic factors and dengue incidences and also determine the time lags of climatic factors in impacting the dengue spread.

Bongwon Suh et al. [8] examined the features which may affect retweetability of tweets. Content and contextual features were gathered from 74M Tweets and this data set was used to identify factors that were significantly associated with retweet rate. He built a predictive retweet model. 10,000 tweets were used to perform exploratory data analysis using Principal Components Analysis (PCA) and Generalized Linear Modelling (GLM). The result showed that 21.1% of tweets have at least one URL in their text. On the other hand we find 28.4%have URLs in them. 10.1% of tweets have atleast one hash tag in their text while 20.8%of retweets contain hash tag. GLM analysis shows that these two features have a very strong relationship with retweet rate. Twitter users who created there accounts more than 300 ago shows a retweet rate higher than the average. Twitter users who created their accounts very recently (<30 days) results in a slight u shaped curve.

DursunDelen et al. [9] applied popular machine learning techniques on large variety of predictive factors by examining the healthcare coverage of individuals. Twenty three variables and 193,373 records were used from the2004 behavioral risk factor surveillance system survey data for this study. Two types of machine learning algorithms were used. They are artificial neural networks and decision trees. Popular artificial neural network architecture called multilayer perceptron is used. Results proved that MLP performs better than other ANN architectures such as RBF, RNN and SOM.ANN model was able to classify those with and without healthcare coverage with an overall accuracy rate of 78.45%.The model classified those with coverage having 80.05% accuracy than those without coverage having 76.86% accuracy. This model was more effective. The decision tree model was 74.11% accurate in classifying with and without healthcare coverage. The decision tree has superior performance in classifying those without healthcare coverage having 75.51% accuracy than those with healthcare coverage having 72.71% accuracy. Income, employment status, education and marital status were the most important predictive features that came out. This study identified the factors that can be used to accurately classify those with and without healthcare coverage.

Fréderic Godin et al. [10] proposed a novel method for unsupervised and content based hash tag recommendation for tweets. Their approach relies on Latent Dirichlet Allocation (LDA) for modelling the underlying topic assignment of language classified tweets. 18 million tweets were collected in which 77% of hash tags were used only once and 94% were not used more than five times. The advantage of our approach is that it can recommend hash tags for tweets in a fully unsupervised manner.

Shuang Yang et al. [11] considered the problem of high precision topic modeling of tweets in real-time as they are flowing through the network. A spectrum of topic modeling techniques that contribute to a deployed system was presented. They have proposed a unique collection of topic modeling techniques that effectively helped us to address the challenges in implementing the system and satisfying the quality requirement.

Ilkyu Ha et al. [12] extracted the representative literature related to Twitter and trends were investigated in Twitter research by using the Systematic Literature Review Method (SLR) method. Total of 12,319 published papers were extracted from the 5 most recommendable public resource sites. 106 papers were finally selected for detail analysis by the proposed SLR.

Table 2. Algorithms and datasets used in the reference papers

Paper Reference	Algorithm used	Data Set used
Ref[1]	Random Forest is used	Data of Twitter and Yelp is used. Yelp data set consist of the data set created from Yelp Data Challenge[13]
Ref[2]	Naive Bayes, Maximum entropy classification and Support vector machine are used.	Cornell movie review data set is taken.
Ref[3]	SVM, C4.5, K-	Dataset is based

	Nearest Neighbor are used	on Web Directories which was used in prior studies. Dataset is from yahoo, ODP and Hoover’s Online Company Database.
Ref[4]	Rocchio’s algorithm, K-Nearest Neighbor, Naïve Bayes, Decision Tree, Decision Rule, SVM, Neural Network, LLSF, Voting, Associative Classifier, Centroid based classifier and Additional Classifier are used.	Data collection set is not specified.
Ref[5]	Expectation-Maximization and Naïve Bayes classifier are used.	The 20 Newsgroups data set containing 20017 articles
Ref[6]	Support Vector Machine is used	Two data sets are used: the first one is “ ModApte” corpus of 9603 training documents and second one is “Ohsumed” corpus
Ref[7]	Support Vector Machine , Genetic Algorithm are used.	Datasets are used from Environmental Health Institute (EHI) of the National Environment Agency (NEA) of Singapore. Climatic data is provided by Meteorological Services Department (MSD) of NEA.
Ref[8]	Principal Components Analysis (PCA) and Generalized Linear Modelling(GLM) are used.	Sample of public tweets are collected from twitter’s API from 18th January,2010 to 8th march 2010.
Ref[9]	Artificial Neural Networks and decision trees are used.	Data was collected from Behavioral Risk Surveillance System 2004 Survey
Ref[10]	Expectation-Maximization and Naïve Bayes method are used. Latent Dirichlet Allocation (LDA) model is used.	18 million tweets are collected from twitter API
Ref[11]	Topic Modelling Techniques such as unsupervised clustering algorithms, information filtering approaches or weakly supervised models are used. two-stage training algorithm and a close-loop inference mechanism are also used.	Twitter data
Ref[12]	Systematic Literature Review Method (SLR) method is used.	12,319 published papers were used from the 5 most recommendable and famous public resource sites

3. CONCLUSION

After reviewing all mentioned papers we came to know that text mining is an important part in which unstructured data can be used for identifying user potentials and interests. We are motivated to work in this area to classify short messages. We selected some twitter text classification related papers from reputed sources (IEEE,Springer etc.). There are mainly two types of work which are focused in the literature: 1. labeled text classification 2.unlabeled text classification. In the case of labelled text classification we use supervised machine learning algorithms to train our classifiers. While in case of unlabeled text classification we use unsupervised machine learning algorithm to train our classifier. After studying all the research that is done on twitter and text classification we decided that we will continue our work in this field. We came to the conclusion that continuing the work on famous microblogging site i.e. twitters is very interesting and offer lot of advantages to the public. Lot of meaningful information can be extracted from the high unstructured twitter data. This meaningful information can also be used to find out the trends in twitter research. Data is categorized into various classes for text classification. Various machine language algorithms are used in the classes to find the outcomes. There are some labeled documents and unlabeled documents and model can be designed from the labeled documents. There are different algorithms of data mining that can be used. Each algorithm classifies data in different manner. Various algorithms can be compared with each other on the basis of their accuracy. They can be compared with each other on the basis of correctly classified instances and incorrectly classified instances of the class. Classification is a very challenging phenomenon nowadays and plays a vital role in research. Text classification uses terms as the features of the class. Useful and relevant features can be find out from the various phenomenon like stemmers, stopwords, TF-IDF score etc. So in future we will be doing text classification and performance evaluation will be done from the relevant features of the classes.

4. FUTURE WORK

In future work can be done to test and validate the text classification algorithms and to study the roles of features in text classification.

5. REFERENCES

[1] Abdulkareem Alsudais, Gondy Leroy, Anthony Corso, 2014,"We know where are you tweeting from:Assigning a Type of Place To Tweets Using Natural Language Processing",IEEE International Congress on Big Data.

[2] Bo Pang, Lillian Lee, Shivakumar Vaithyanathan, 2002, "Thumb up Sentiments Classification Using Machine Learning Techniques*",* Proceedings of EMNLP.

[3] Evgeniy Gabrilovich , ShaulMarkovitch , 2004, "Text Categorization with Many Redundant Features: Using Aggressive Feature Selection to Make SVMs Competitive with C4.5*",* ICML.

[4] VandanaKorde, C NamrataMahender, March 2012, "Text Classification And Classifiers: A Survey", International Journal of Artificial Intelligence & Applications(IJAIA),Vol 3, No 2,.

[5] Kamal Nigam, Andrew,Kachites, Mccallum, Sebastian Thrun, Tom Mitchell,"Text Classification from Labeled and Unlabeled Documents Using EM", Kluwer Academic Publishers, Boston. Manufactured in Netherlands.

[6] Thorsten Joachims," Text CategorizationWith Support Vector Machines: Learning With Many Relevant Features"

[7] Xiuju Fu, Christina Liew, Harold Soh, Gary Lee, Terence Hung, Lee-Ching Ng, 2007,"Time Series Infectious Disease Data Analysis Using SVM AND Genetic Algorithm",IEEE.

[8] Danah Boyd, Scott Golder, Gilad Lotan, 2010," Tweet, Tweet, Retweet: Conversational Aspects of Retweetingon Twitter",IEEE.

[9] DursunDelen, Christie Fuller, Charles McCann, Deepa Ray, 2007," Analysis of healthcare coverage: A Data Mining Approach",Expert systems with applications

[10] Fréderic Godin,Viktor Slavkovikj, Wesley De Neve," UsingTopic Models for Twitter Hashtag Recommendation", International World Wide Web Conference committee (IC3W2).

[11] Shuang Yang, Alek Kolcz, Andy Schlaikjer, Pankaj Gupta, "Large-Scale High-Precision Topic Modeling on Twitter", in the proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data

[12] Ilkyu Ha, Hohwan Park, Chonggum Kim," Analysis of Twitter Research Trends based on SLR",Advanced Communication Technology(ICACT), 2014 16th International Conference