Study of supervised machine learning approaches for sentiment analysis

Sangharshjit S. Kamble[1], Prof. A. R. Itkikar[2]

[1] Student Master of Engineering (CE), Sipna COET, Amravati
[2] Professor, Dept of Computer Science & Engg, Sipna Coet, Amravati, Maharashtra

---***---

Abstract - *The field of sentiment analysis is the more attractive field of data mining and data analysis. Sentiment analysis is use in a various departments like ecommerce sites, government organization and other department. The sentiment analysis are use to identify opinion of the people about any aspect or product. There are various algorithm used for sentiment analysis. These algorithms are broadly categories into two types i.e. Lexicon based approach and Machine learning approach. Each approach has some advantages over another algorithm. In this paper, we study the supervised machine learning approach and its algorithms.*

***Key Words*: NB, SVM, ME, Sentiment, opinion**

1. INTRODUCTION

Sentiment Analysis is the study of people opinions, attitudes and emotions toward an entity. The entities are the event, aspect or any other that concerned with people opinion. The two domains SA and OM are mutual properties. They express some type of mutual meaning. Some of the researcher stated that Opinion Mining is using to extracts the people's opinion about the any entity or aspect. On the other side Sentiment Analysis identifies the sentiment that expressed into the text then analyzes it. So, the target of SA is to find opinions, identify the sentiments they express, and then classify their polarity i.e. positive, negative or neutral. Sentiment Analysis is the classification process. The classification process are divided into the several part depends on there working these are Document -Level, sentence-level and the aspect level

1.1 Sentiment Analysis

The sentiment analysis is the process to identify the opinion or the people's emotion about any entity. Sentiment analysis provides us a way to identify the people's opinion about the entity. There are two approaches to identify the sentiment of the people about the entity. The sentiment analysis processes are use lexical base approach and machine learning approach. In the lexical base approach the classification of sentiment is based on the analysis of individual words and/or phrases; in lexical analysis the dictionary are used. Emotional dictionaries are frequently used: lexical items are searched into the dictionary. The dictionary contains sentiment value of each word. The word that is store in emotional dictionary has some weighted value assign with it. The texts that analyze for sentiment are compared with the stored value. After the comparing weighted value explore the polarity of the texts. Polarity of the text is positive, negative or neutral. On the other side the working machine learning approach are different than the lexical base approach. In machine learning approach the classifier are divided into the two parts. Before the text are analyze for the sentiment first we have to train the classifier. On training of classifier it labeled with it polarity value. The labeled data are store into the train data set. In machine learning approach there are two data set are used i.e. train data set and test data set. The train data set are trained by the algorithm before the sentiment analysis. The other test data set are use to check the polarity of the text. Each approach has its advantages and disadvantages. When using the lexical approach, there is no need for labeled data and the procedure classifies the train data and the decisions taken by the classifier. On the other side when using the machine learning there is no need of emotional dictionary, also no need to check the polarity of each word.

1.2 Classification Process

The classification process plays the important role in sentiment analysis. For the classification there are two approach are use lexicon based approach and machine learning approach.

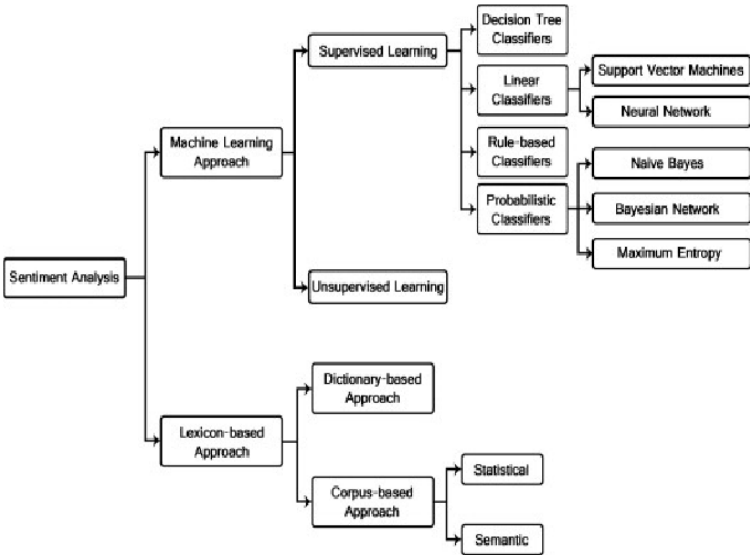

Fig.1. Classification Process

In the figure 1 show the classification process of sentiment analysis. The machine learning approach are further divided into two types supervised learning and unsupervised learning. In this paper we focused on the supervised machine learning approaches and these approaches are Naïve bayes (NB), support vector machine (SVM) and maximum entropy (ME).

1.3 Machine learning

Machine learning is the science of getting computers to act without being explicitly programmed. The machine learning approaches are different than the traditional computing, algorithms are use to train the data explicitly to problem solve of sentiment analysis. Machine learning algorithms instead allow for computers to train on data inputs and use statistical analysis in order to output values. By using train data set, machine learning allow computer in generate models from the stored data in order to automate the system. It also allows system to decision-making processes based on data inputs. In machine learning approach two set of data are required i.e. train data set and test data set. Two of the most widely use machine learning methods is supervised machine learning which trains algorithms based on example input and output data that is labeled by humans, and unsupervised learning which provides the algorithm with no labeled data in order to allow it to find structure within its input data. In the following subsection explore the detail of each approach.

1.4 Supervised learning

The supervised learning methods depend on the existence of labeled training documents. In supervised machine learning, there are two data set are used. One data set known as train data set are used to train the data on selected classifier. In this train data set the computer provide with the inputs that are pair with input and with their desired output. The main motive of this method is for the algorithm to be able to "learn" by comparing its actual output with the "taught" outputs to find polarity. Supervised learning uses patterns approach to predict label values on additional unlabeled data.

For example, suppose you had a basket and filled it with different kinds of fruits. Your task is to arrange them into groups. We have four types of fruits. They are apple, banana, grape, cherries. You already learn from your previous work about the physical characters of fruits. So arranging the same type of fruits at one place is easy now. In data mining and sentiment analysis the earlier work is called as training the data. Now in the train data set you have physical properties of fruits i.e. size, color, shape, and name. Suppose you taken a new fruits from the basket then you will see the size, color, and shape of that particular fruit. If the size is big, color is red, the shape is rounded shape with a depression at the top, you will confirm the fruits name as apple and you will put in apple group, likewise for other fruits also. If you learn the thing before from training data and then applying that knowledge to the test data, this type of learning is called supervised learning. Images as fish and unlabeled ocean images as water. The same process will work on sentiment analysis; in trained data set we provide the polarity of word i.e. positive, negative and neutral.

Following are some of the supervised machine learning approaches that are mostly used.

1.4.1 Naïve bayes

Naïve bayes algorithm works on probabilistic approach. These algorithms are use in supervised machine learning. The main purpose of naïve bayes algorithm is to train classifier. Further this train classifier are use to classify test data. The working principles of this algorithm are common to processes. These working principles are: the classifiers that are train to classification problem simply assume that the value of a one particular feature is to be independent of the value of any other feature, given the class variable. The Naive Bayes algorithm is the simplest and most common algorithm used for train classifier. Naive Bayes classification model calculate the probability of a class, that base on how much time words are appear into the documents. The model works with the BOWs feature extraction which ignores the position of the word in the document. It uses Bayes Theorem to predict the probability that a given feature set belongs to a particular label.

$$p\,(\text{Label/feature}) = \frac{p(\text{Label})*p(\text{feature/Label})}{p(\text{feature})}$$

P (Label) is used to show the prior probability of a label P (features/label) is the prior probability that a given feature set is being classified as a label.

1.4.2 Support Vector Machine

The support vector machine uses the linier properties of probability. The main principle of SVMs is to determine linear separators in the search space which can best separate the different classes. In Fig.2 there are 2 classes x, o and there are 3 hyperplanes A, B and C. Hyperplane A provides the best separation between the classes, because the normal distance of any of the data points is the largest, so it represents the maximum margin of separation.

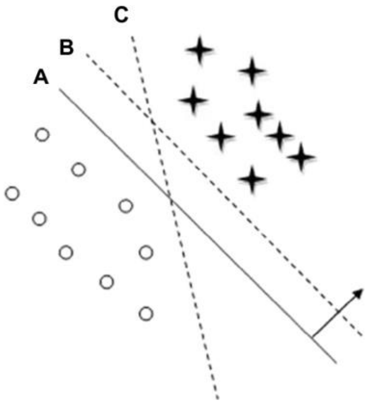

Fig- 2: Using support vector machine on a classification problem.

Text data are ideally suited for SVM classification because of the sparse nature of text, in which few features are irrelevant, but they tend to be correlated with one another and generally organized into linearly separable categories. SVM can construct a nonlinear decision surface in the original feature space by mapping the data instances non-linearly to an inner product space where the classes can be separated linearly with a hyper plane.

1.4.3 Maximum Entropy

The Maxent Classifier (known as a conditional exponential classifier) converts labeled feature sets to vectors using encoding. This encoded vector is then used to calculate weights for each feature that can then be combined to determine the most likely label for a feature set. This classifier is parameterized by a set of X{weights}, which is used to combine the joint features that are generated from a feature-set by an X{encoding}. In particular, the encoding maps each C {(feature set, label)} pair to a vector. The probability of each label is then computed using the following equation:

$$P(fs/label) = \text{¼}\frac{\text{Dot prod (weights; encode (fs, label))}}{\text{Sum(dotprod(weights;encode(fs,l))forlinlabels)}}$$

2. Result

The supervised machine learning approach are use to identify the sentiment about the text or peoples opinion. There are various algorithm are use in supervised machine learning approach. These algorithms are naïve bayes, support vector machine and maximum entropy. These algorithms are the probabilistic algorithm that uses the mathematical expression to classify the sentiment about the texts. Accuracy of each algorithm is varying from aspect. Apart from various algorithm naïve bayes algorithm provide good accuracy in sentiment analysis. Other two algorithm i.e. support vector machine and maximum entropy also provide a good result on some data.

3. Conclusion

In this review article I am try to cover these three supervised machine learning algorithm. In this study we learn what exactly the sentiment analysis is and the various approaches that are use for sentiment analysis. These two are unsupervised and supervised machine learning approach. Further supervised machine learning is use two types of data set i.e. Train data set and test data set.

REFERENCES

[1] Blinov P. D., Klekovkina M. V., Kotelnikov E. V., Pestov O. A. "Research of lexical approach and machine learning methods for sentiment analysis ".

[2] B. Pang, L. Lee, and S. Vaithyanathan, Thumbs up? Sentiment Classification using Machine Learning Techniques, Proc. Of the Conf. on Empirical Methods in Natural Language Processing (EMNLP)

[3] Doaa Mohey El-Din Mohamed Hussein" A survey on sentiment analysis challenges", In Ain Shams Engineering Journal (2014) 5, April 2016

[4] David Osimo and Francesco Mureddu "Research Challenge on Opinion Mining and Sentiment Analysis.

[5] Kumar Ravi, Vadlamani Ravi "A survey on opinion mining and sentiment analysis: Tasks, approaches and applications" In Knowledge-Based Systems 89 (2015)

[6] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, Manfred Stede "Lexicon-BasedMethods for Sentiment Analysis"

[7] Nipun Mehra , Shashikant Khandelwal , Priyank Patel "Sentiment Identification Using Maximum Entropy Analysis of Movie Reviews"

[8] S. M. Vohra, J. B. Teraiya "A Comparative Study of Sentiment Analysis Techniques" In Journal of Information, Knowledge and Research in Computer Engineering

[9] Walaa Medhat , Ahmed Hassan , Hoda Korashy "Sentiment analysis algorithms and applications: A survey" In Ain Shams Engineering Journal (2014) 5, 1093–1113