Survey on Sentiment Analysis and its Classification Technique

Amit K. Burde
UG Student
Computer Engineering SPPU
Pune University
Maharashtra, India

Tushar G. Barage
UG Student
Computer Engineering SPPU
Pune University
Maharashtra, India

Vikram B. Shevate
UG Student
Computer Engineering SPPU
Pune University
Maharashtra, India

Preeti Suryawanshi
UG Student
Computer Engineering SPPU
Pune University
Maharashtra, India

Viresh Chapate
Assistant Professor
Computer Engineering SPPU
Pune University
Maharashtra, India

ABSTRACT
In today's world everyone is trying different product and services. They are always commenting on such things on microblogging sites like Twitter and Facebook. Sentiment analysis also known as opinion mining used for finding the polarity of people's opinion,thoughts, reviews or evaluations posted on the internet. Now these opinions differ from person to person but they either have positive, negative or neutral characteristics. Field of sentiment analysis is gaining importance and is used in many fields. So finding proper attitude of one's opinion is very important. Different algorithms are used in sentiment analysis based on the inputs that have been taken. Proper understanding of sentiment analysis and its algorithm will make it possible to analyze the opinion and produce the accurate result of them in terms of positivity and negativity

Keywords
Sentiment Analysis, SentimentClassification techniques, Text Extraction, Naïve Bayes, Bayesian Network, Random Forest, Support Vector Machine, Artificial Neural Network.

1. INTRODUCTION

1.1 Sentiment Analyses
In simple words, process of Sentiment analysis is the task of classifying whether the sentiments, thoughts, reviews posted in a text is showing positivity or negativity about product, service or person. Sentiment Analyses focuses on dividingtextby their opinion and emotions expressed by the person. Finding atext's polarity as positive or negative is a two-class problem.This is known as sentiment orientation analysis in text classification process. [1] Text classification has been being very helpful. With the rapid development ofthe internet, web and smart devices users are always posting their opinions on the internet. feedback and comments based on those gives the marketing people improve their product and services. So Sentiment Analysis is being used as a great marketing tool. Importance of sentiment analysis is increasing greatly.

1.2 Example on polarity/attitude of sentiment
a. I'm feeling so well today. – Positive

b. I love the way Arsenal players play football – Positive

c. I don't like sedans – Negative

d. I think windows 10 will be a good OS – Positive

1.3 Application of Sentiment Analysis
- Analyzing customer feedback can make it possible to make the product better
- Finding Happiness meter of the customers
- Finding recommended products, features from customer
- Stocks can rise or fall in a split Seconds. Sentiment analysis can predict the stock moods

Above things make sentiment analysis a great market tool

1.4 Sentiment Analysis Phases

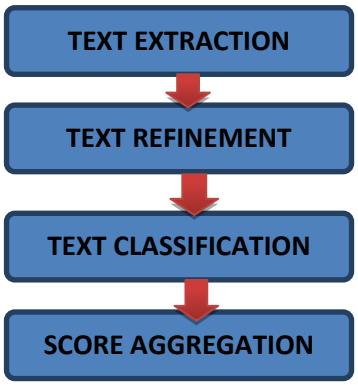

Fig 1: Phases Of Sentiment Analysis

1.4.1 Text Extraction
[2]For making food there is need of vegetables, herbs etc. Just like that for text analysis some kindof textual data is needed. So first there is a need to determine the source for extracting text. This data can be taken from twitter, surveys and other databases. After that will this text can be loaded into processing system. After that the removal of unwanted words or meaningless words should be done. Also there can be no of emoticons in the text, so there is a need to manage that too.

1.4.2 Text Refinement
After extracting the desired text, parsing of the contents from that text will be done. Words from text are divided into segments. Terms present in the text sentences are identified

1.4.3 Text Classification
Text might suggest positive, negative even sometimes neutral attitude.

1.4.4 Score Aggregation

After determining the phrases bearing the sentiments of person during text classification collection of these phrases as well as scores are assigned based on positive and negative opinion. Based on scores attitude of the sentiment of text can be found.

2. CLASSIFICATION

Classification is a phase in sentimental analysis which is used to describe the process in which predictionof the sentiment or emotions of the response is determined.Identifying the sentimental response of the document or emotion includes the different classes.There are different techniques of classification are out there which can be used to build classifiers.

2.1 Naïve Bayes Classifier

[1] Naïve Bayes classification is bases on Bayer'stheorem.Naïve Bayes' classifier is one of the oldest, most used and successful classifier. It is one of the best classification as it requires less time & memory space in the system. [3] Naive Bayes technique is so simple that one not being skillful in classifier technology will be able to understand it easily and implement it. It's based on making unrealistic independence assumption of variable. In this classifierit is considered that the classes of the classification & documents are totally independent from each other. Naive Bayes' classifier provides unreasonable efficient result. There are some results which shows that even when there werevery low chances of getting efficient classification but still it gives the proper & perfect result.

It is generally assumed & used in real time processes also for getting correct result of the decision. It provides a token or code to energy word in the statement & according to that token all sorting & decision making process are done. It creates a set which contain all the subsets & in those subsets, these text & words are stored according to their token classification. It is used in various real world problems like sentimental analysis, spam detection in e-mail, auto grouping of e-mails, sorting of e-mails according to user's priority. [4] The best advantage of the Naive Bayes' is it required low memory & less time for complete execution. This classifier should be used when the training time is one of the most important factor for the system. When limited amount of resourcesareavailable with the CPU than this classifier is used or is advised to use it.

[5]Song Qing, Liu Xisheng, Yuan Hui and Qiu Chen proposed a system of Droplet Fingerprint Recognition by applying the NaiveBayes classifier. Their system gave them 98.765 % rate of recognition with the help of Droplet. Sichuan, China, Wuming Pan, Haiming.

[6]Li, Yang Xu proposed a system for fuzzy clustering using Naïve Bayes classifier. They showed that fuzzy Naïve Bayes classifier can be effective solution for dealing with the problem of classification where there are continuous variable.

2.1.1 Naive Bayes is further classified:

2.1.1.1 Multinomial NaiveBayes:

When a word is occurred for the multiple times in text classification problems.
Example - Top classification

2.1.1.2 Binaeized multinomial Naive Bayes:

It does not consider that how many times a particular word is used in the statement but it considers by what sense it has been used.

Example - sentimental analysis

2.1.1.3 Bernouli Naive Bayes:

It is used when some word or content is missing & which is effecting the entire statement or the problem.
Example: It is used to check spams.

2.1.2 Bayes' Theorem

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)} \qquad (1)$$

Above P(X|Y) is posterior. P(Y|X) is prior P(X) is likelihood.

P(Y) is evidence.

2.1.2 Advantages of Naïve Bayes Classification

- Gender classification
- Document classification
- Spam filtering

2.1.3 Advantages of Naïve Bayes Classification

- It's very simple to implement and requires very less effort too.
- It requires very little training data.
- Though there can be variation most of the times it gives good results.

2.1.4 Disadvantages Naïve Bayes Classification

- Strong feature independence assumptions are made which causes loss of accuracy.
- Dependency among some variable cannot be modeled in Naïve Bayes classification.

2.2 Bayesian Network

[2] In Naïve Bayes strong feature independence assumptions are made, but in Bayesian network one assumes that all features are dependent. With the help of directed acyclic graph (DAG) Bayesian networkcan be represented visually. Below is an example of Bayesian network. A, B, C and D are variables in the network which are represented by circles. Relationship between them is represented by arcs, arrows shows the dependency between these nodes. This allows computation of probabilities of every node in network. It is considered one of the most complete model because of the ability to representing the variables their dependency and joint probability distribution (JPD).

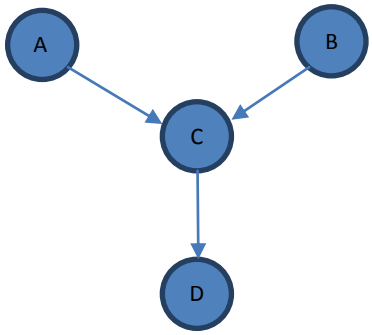

Fig2: Bayesian Network

[7]Structure evolution of dynamic Bayesian network for traffic accident detection Ju-Won Hwang, Young-Seol Lee, Sung-Bae Cho used dynamic Bayesian network for detecting traffic accident. They suggested that dynamic Bayesian network overcomes the problem of typicalBayesian network where increased no of nodes makes it harder to define parameter and structure.[8]Bo Chen, Qin Liao, Zhonghua Tang wanted to make their Bayesian network classifier more accurate and less time consuming. Since increased nodes makes typical Bayesian network times consuming and less accurate so they used clustering in Bayesian network. They divided the Bayesian network structure by using hierarchical clustering and performed searching in these smaller state spaces.

2.2.1 Application of Bayesian Network

- Computational biology
- Gene and protein analysis
- Classifying documents
- Retrieval of information
- Risk analysis
- Weather prediction

2.2.2 The advantages of Bayesian Network

- Representationof Bayesian network can be done visually.
- Easy to understand relationship between various nodes.
- Bayesian networks can handle situations where the data set is incomplete since the model accounts for dependencies between all variables.
- Bayesian networks can map scenarios where it is not feasible/practical to measure all variables due to system constraints (costs, not enough sensors, etc.)
- Help to model noisy systems.
- Can be used for any system model - from all known parameters to no known parameters.

2.2.3 The disadvantages of Bayesian Networks

- All branches must be calculated in order to calculate the probability of any one branch.
- Accuracy of result depends on the model itself and prior beliefs.
- In Bayesian Network calculation are complex and very costly.

2.3 Random Forest Classification:

Random forest classification is ensemble method it uses multiple learning approach of algorithm to obtain the predictive result. It is based as decision tree aggregation. It takes the multi attitude decision tree as an output. The performance graph for this classifier is always in increasing order it never decrease. With the better performance it also provides excellent accuracy. This classifier is widely used in various application such as Biometrical application sexually explicit content detection. There can be multiple data sets of random forest classifier. As they vary from application to application.

Random forest provides&support parallelexecutions& multi core algorithms. So it helps simultaneousrunning of different trees at a time. The time required for training classifiers is recurrent learning with every normal dataset.

[9] Farooq, F.,Aarhus, Denmark, Kidmose, P. used random forest for classification of P300. 2 datasets were used they were BCI competition dataset and image driven paradigm dataset. Their method showed great accuracy

[10] Gislason, P.O,Reykjavik, Benediktsson, J.A., Sveinsson, J.R.used random forest method for classifying the multisource data. They compared their results obtained using the method random forest to the results obtained by using the method of bagging and boosting.

2.3.1 Advantages of Random forest

- It is easy to implement& to understand.
- It supports simultaneous & scalable classification &it's also fast.
- It automatically generates the variable taken for classification in aclass according to their importance & relevance.

2.3.2 Disadvantages of Random forest

- The major disadvantage in Random forest classifier isit easily over fit sothere is a need to reduce the number of trees & as well as need to reduce the vague links which are present.
- Large no of trees can make the performance of randomforestslower.

2.4 Support Vector Machine (SVM)

SVM is a non-probabilistic binary linear classifier because SVM classifies its input into 2 possible classes. SVM can also be used to perform a non-linear classification with the help of kernel trick. implicitly mapping their inputs into high-dimensional feature spaces.Suppose there are Black and white circles and if there is a need to separate them. One can do this by placing many straight lines. But only one with the Maximum Margin should be selected since it givesthe optimal result. Finding the optimal hyperplane is the main objective of SVM.

But if onewants to place a curved line than the Straight one then it is not possible. For this type of non-linear classification, there is a technique called the Kernel Trick. This allows operating of inputs on the high dimensional spaces.

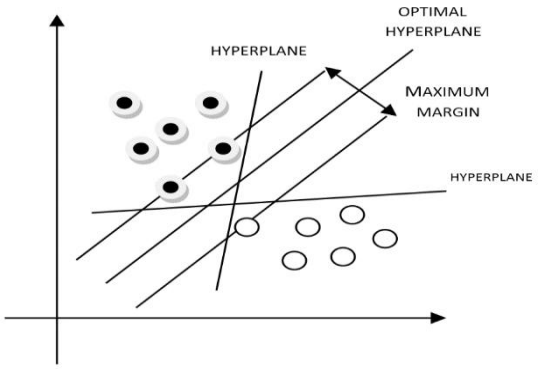

Fig3: Linear Classification

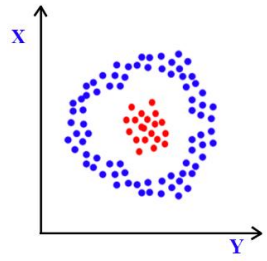

Fig4: Before Classification of Non-linear Data

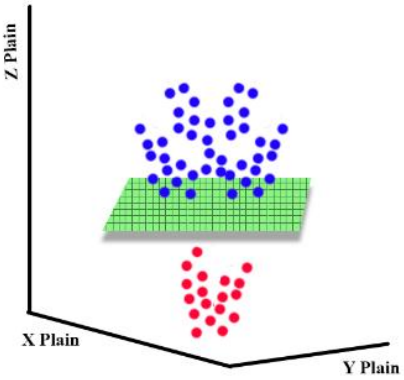

Fig5: Circles in Hyperplane

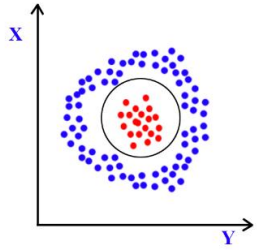

Fig6: After Classification of Non-linear Data

[11] For evaluation of credit risk Yongqiao Wang. Shouyang Wang, Lai, K.K. proposed A new fuzzy support vector machine. They proposed to use fuzzy support vector machine to differentiate the good creditors and the bad creditors. They showed that instead of two group classification they wanted to use quadratic problem, and it stated that it can be more powerful than the standard support vector machine. But also for that to happen a proper and appropriate kernel must be chosen.

[12] Lei Li,Zhi-ping Gao, Wen-yan Ding proposed a network intrusion system which exploited the Fuzzy Multi-class Support Vector Machine. They used it because Fuzzy multi class support vector machine can work better than standard support vector machine because it can overcome the problem of noise non linearity and uncertainty. Their method had better detection accuracy as well as the training time was also significantly reduced.

$$Lp = \frac{1}{2} \| w^2 \| - \sum_{i=1}^{t} \alpha i yi \, (w.xi + b) + \sum_{i=1}^{t} \alpha i \quad (2)$$

With respect to **w** and *b* in this function, in this t is the no of training examples. i = 1…., t, these are non-negative values. *αi* are the multipliers of Lagrange, *LP is* Lagrangian, w and b defines the hyperplane.

2.4.1 Applications of Support Vector Machine

- Using SVM recognition of the hand written texts and patterns is possible.
- In Medical field one can use SVM to identify to classify the proteins.
- Classification of images can be done using SVM
- Hypertext characterization.
- Classifying the reviews based on their quality.

2.4.2 Advantages of Support Vector Machine

- High accuracy.
- On Text data SVM works better.
- Have better memory efficiency.

2.4.3 Disadvantages of Support Vector Machine

- Computational Efficiency.
- Have difficulties in dealing with multiclass problem.
- Using kernel trick for nonlinear classification may make SVM computational large.

2.5 Artificial Neural Network(ANN)

Artificial Neural Network are based on biological neural network of brain. They are doing parallel operations on the nonlinear data much like the human neuron system. The no of neural nodes in Artificial neural network is very small as compared to the no of neurons in Biological Neural network also they are easy to understand. There are 3 layers in Artificial neural network input layer, hidden layer and output layer. Input layer takes some form of hidden pattern as input and then forwards it to the hidden layer. Hidden layer is where actual processing happens. Many hidden layer can work together to process the inputs. Usingfeed-forward technique hidden layer is constantly forwarding its output to the next layer of hidden layer which eventually will transfer that output to the output layer. Where in feed forward every layer is connected to next layer in recurrence network are connected to previous layers which makes it like feedback system.

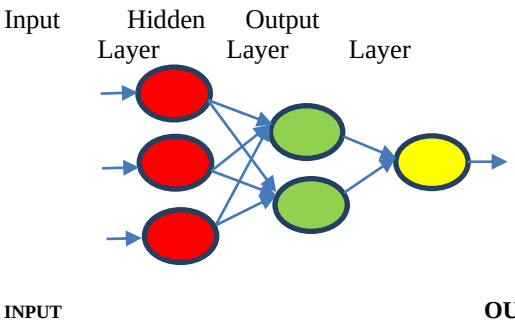

Fig 7: Neural Network Structure

[13] Mahmon, N.A., MARA, Shah Alam, Ya'acob, proposed the method in which they used Artificial Neural Network for classifying satellite images. They wanted to try out different algorithms for this purpose and that included using of back-

propagation K means Algorithm. To analyze the performance of classification of satellite images values like kappa coefficient were calculated.

2.5.1 Machine Learning in ANNs

Most interesting thing about Neural Network is their ability of learning from examples.

2.5 .1.1 Supervised Learning –

In supervised learning inputs and the outputs which are given one can use neural network to model the relationship between these twos. So neural network after processing compares the result with the expected output and does the needed adjustments according to error.

2.5 .1.2 Unsupervised Learning

No desired output is provided in this type of learning method. So there is a need to do grouping of given data on the basis of some trait or characteristics, this is called clustering.

2.5 .1.3 Reinforcement Learning

This type of learning is based on observation. The ANN tries to make a decision by the observation of surrounding environment. For negative observation, network makes necessary adjustment so that it can make an alternative on next step.

2.5 .2 Applications of Neural Networks

They can perform tasks that are easy for a human but difficult for a machine

- In aerospace technology as auto piloting the aircrafts, fault detection in airplanes.
- In military field for target tracking, missile guide system.
- For financial analysis.
- In medical field for analyzing cancer cells, EFG, ECG.
- Can be used in speech recognition and classification. One can use them for text to speech conversion.
- For Pattern, facial, character recognition, etc.
- In signal processing.

2.5 .3Advantages of Neural Networks

- Useful in modeling nonlinear data
- Higher accuracy
- Easy Understanding and implementation

2.5.4Disadvantages of Neural Networks

- There is a lack of transparency in neural networks. Neural networks are like a black box.
- Training time is more.
- Multilayer neural networks are usually hard to train, and require tuning lots of parameters.

3 CONCLUSION

Sentiment analysis involves the practice of natural language processing and text analyzing over textual data, to find the attitude of sentiment. With the rise of the internet, smart phones and social networking sites, people are showing their responses over products. Analyzation of that response can lead to better market analysis. Sentiment analysis and opinion mining have become one of the best marketing tool. Sentiment analysis can make possible to produce the best product possible by making the analysis of the response of people

In this survey, comparison of some well-practiced classifiers was done on the basis of accuracy, complexity, performance, required training data, etc. Result of survey tells that every algorithm has its own characteristics which makes it better and worse than the one we are comparing. Naïve Bayes is the simplest of them all requires very less training data but has variable accuracy. Bayesian network can be used to model most systems and one can represent it visually using it too. Finding of probability of every node in the network is possible. But computation in Bayesian network is very expensive. Random forest is very fast and scalable too. But overfitting can be a problem to handle so that there might be a need to remove some links. Random forest can rank their variables by their importance.Support vector machine if used appropriate kernel can be a powerful classification technique which will give you higher accuracy and very good memory efficiency. While being easy understanding as well as easy, implementation is a good trait of Artificial Neural Network is lack of transparency can be a problem.

So,if you want simplicity and there is very less training data as well as limited resources you must use Naïve Bayes without any hesitation. Because any novice can be able to understand and implement Naïve Bayes Classifier. If you want performance, accuracy, want to rank the variable by their importance Random Forest is a must which even outperforms support vector machine too. With the proper selection of classification techniques, one can get faster and accurate result which will be computationally inexpensive too.

4 ACKNOWLEDGEMENT

To achieve anything in life there is need of support and help from our mentors, teachers and friends. We are very great full for all experts who have contributed and guided us in making this survey possible. This survey would not have been possible without their guidance, expertise and resources. We will like to thank to our family members and friends for encouraging us to be better than our own expectation.

5 REFERENCES

[1]. "Comparative Study of Classification Algorithms used in Sentiment Analysis" Amit Gupte, Sourabh Joshi, Pratik Gadgul, Akshay Kadam.

[2]. "Sentiment analysis algorithms and applications: A survey" Walaa Medhat, Ahmed Hassan, Hoda Korashy.

[3]. " Top 10 algorithms in data mining "Xindong Wu · Vipin Kumar · J. Ross Quinlan · Joydeep Ghosh · Qiang Yang · Hiroshi Motoda · Geoffrey J. McLachlan · Angus Ng · Bing Liu Philip S. Yu· Zhi-Hua Zhou · Michael Steinbach · David J. Hand · Dan Steinberg.

[4]. "Machine Learning Tutorial: The Naive Bayes Text Classifier" Vasilis Vryniotis http://blog.datumbox.com/machine-learning-tutorial-the-naive-bayes-text-classifier

[5]. "Naive Bayes Classifier Applied in Droplet Fingerprint Recognition" Song Qing, Liu Xisheng, Yuan Hui, Qiu Chen.

[6]. "Fuzzy Naive Bayes classifier based on fuzzy clustering " Yongchuan Tang, Sichuan, China Wuming Pan, Haiming Li ; Yang Xu.

[7]. "Structure evolution of dynamic Bayesian network for traffic accident detection" Ju-Won Hwang, Young-Seol Lee, Sung-Bae Cho.

[8]. "A Clustering Based Bayesian Network Classifier" Bo Chen, Guangzhou, Qin Liao, Zhonghua

[9]. "Random forest classification for p300 based brain computer interface applications" Farooq, F.,Aarhus, Denmark, Kidmose, P.

[10]. "Random Forest classification of multisource remote sensing and geographic data" Gislason, P.O, Reykjavik, Iceland, Benediktsson, J.A., Sveinsson, J.R.

[11]. "Fuzzy Multi-class Support Vector Machine Based on Binary Tree in Network Intrusion Detection" Lei Li, Zhi-ping Gao, Wen-yan Ding.

[12]. "A new fuzzy support vector machine to evaluate credit risk" Yongqiao Wang, Shouyang Wang, Lai, K.K

[13]. "A review on classification of satellite image using Artificial Neural Network (ANN)" Mahmon, N.A, Shah Alam, Ya'acob.