Workflow of Machine Learning Based Sentiment Classification: A Review

Poonam Choudhari[1], Dr. S. Veena Dhari[2]
[1] Research Scholar, Dept. of CSE, RNTU, Bhopal (M.P.) India
[2] Associate Professor, Dept. of CSE, RNTU, Bhopal (M.P.) India

ABSTRACT
Sentiment Analysis is the process of understanding the meaning of feelings expressed by an individual online about entities such as products, services, organizations, individuals, issues, events, topics, and interpreting them in positive, negative and neutral classes in a automated way. The paper explains the workflow of Machine learning based approach classification for Sentiment Analysis .It gives an overview of common techniques that are used at different phases of classification with their strength and weakness. This study would help in solving research problems that are encountered in Sentiment Analysis. The understanding of the phases would help in developing efficient classifier and ultimately improving the performance of the Classification algorithm. A study of techniques adopted by researchers at different phases of Sentiment Analysis and brief analysis of study is also presented.

Keywords:-Sentiment Analysis, Feature Vector, Customer Reviews, Text preprocessing, Machine Learning based Classification.

I INTRODUCTION

Sentiment Analysis (also called as Opinion Mining) is the process of classifying whether a piece of writing (online review, tweet, etc.) expressed by opinion writer is positive, negative or neutral. The process is performed by analyzing the different customer online reviews .It is a way of identifying customer attitude, sentiments towards company's product, brand, or service [1]. The online review helps the new customers to take purchase decisions by reviewing the customer feedback reviews. The companies can measure the customer's satisfaction and give a better picture where they stand against their competitors in the market. Since for popular products, a thousand of reviews are available, the analysis cannot be done manually, so an automated way of analyzing the sentiment is required, which is called as Sentiment Analysis.

II LITERATURE SURVEY

A lot of research work has been done in the field of Sentiment Analysis and it is an ongoing research. Pang et al.(2002)[2] used machine learning approach for sentiment classification They experimented with unigrams, bigrams, position based features, POS Based features adjectives, adverb and their combination as features.

They used naïve bayes, SVM and maximum entropy as classification algorithm. Their experimental results showed that unigram features perform best among these features with SVM as classification algorithm with accuracy 82.9%. Pang et al(2008)[3] explained in the paper the research challenges ,applications of sentiment analysis The author provide a brief overview on classification, feature extraction in sentiment analysis, opinion summurization and a list of datasets used by researchers in sentiment analysis. Bing Liu (2012)[4]provides a broad overview of different approaches and techniques in sentiment analysis. The author encourages solving challenges in the field and openly makes resources available for the required work. Narayanan et al.(2013)[5] used text preprocessing technique like use of negation handling, for feature extraction used bag of n-grams and for feature selection used mutual information and observed a significant improvement in accuracy with Naive Bayes classifier. They achieved an accuracy of 88.80% on the popular IMDB movie reviews dataset. Jose et al(2015)[6] present the results for classifying the sentiment of movie reviews which uses a chi-squared feature selection mechanism, machine learning algorithms such as Naive Bayes and Maximum Entropy can achieve competitive accuracy . It analyze accuracy, precision and recall of machine learning classification mechanisms with chi-squared feature selection technique and The method also uses a negation handling as a pre-processing step in order to achieve high accuracy. Kaur et al.(2016)[7] presents an empirical study of efficacy of classifying product review by semantic meaning The authors hereby propose completely different approaches including spelling correction in review text, and then classifying comments employing hybrid algorithm combining Decision Trees and Naive Bayes algorithm. Mubarok et al (2017)[8] experimented on aspect based sentiment classification on product reviews. They used Part-of-Speech (POS) tagging, feature selection using Chi Square, and classification of sentiment polarity of aspects using Naïve Bayes with its highest F1-Measure of 78.12%. The Chi Square also has been proven to speed up the computation time in the classification process of Naïve Bayes although it degraded the system performance.

III WORKFLOW OF SENTIMENT ANALYSIS

The basic steps of sentiment analysis are as follows [9][10]:

(i) Text preprocessing
(ii) Feature vector construction
(iii) Feature Extraction
(iv) Feature Selection
(v) Feature Weighting
(vi) Machine learning based Classification on Feature vector

A brief description of how the different steps contribute in building the sentiment classifier. It also gives a description of techniques used in the phases, with their strength and weakness.

(a) Text preprocessing

In this phase, the text is cleaned by removing the irrelevant things from the text and makes it ready for its next phase.

The texts are pre-processed by following methods:

(i) **Tokenization**

Text data is converted to a block of characters called tokens. The tokens/words which will be used as feature vector are used for further processing.

(ii) **Strength of Tokenization**

- Tokenization is the first major step in text pre-processing which convert the raw data into identifiable data known as tokens.

(iii) **Weakness of Tokenization**

- Tokenization is one of the crucial steps in pre-processing of text. The major issue here is to recognize the correct tokens. For e.g. splitting on white space can split the word which should be regarded as single token like Los Angles.

(iv) **Removal of Stop Words**

Stop words are the words that occur so often in the text, but support no information for classification. For e.g. In English language 'a', 'an', 'the' are considered as Stop Words.

(v) **Strength of Stop Words**

- Stop words removal helps to reduce the dimensionality of the data space. Thus learning becomes faster classification can be more accurate as the noise in the form of stop words are removed.

(vi) **Weakness of Stop Words**

- The list of stop words may vary according to domain/language, as some of the words may have relevance in some domain than in other domain. So there is no definite stop word list which works in all type of domain applications.

(vii) **Stemming**

The tokens are reduced to their stem. In this technique, the suffix /prefix are removed from the words for e.g. 'watched' becomes 'watch'.

(viii) **Strength of Stemming**

- Stemming optimize the performance of algorithm by reducing the size of data space.

(ix) **Weakness of Stemming**

- Stemming should be done carefully. If not, it may lead to over stemming or under stemming. Over stemming means two words with different stems are rooted to same word which may give false positive. Under stemming means when two words that should be stemmed to same word are not which may give false negative?

(b) Feature vector construction

When the cleaned text is obtained as output from pre-processing phase, the feature vector is constructed which will be given as input to next phase i.e. classification. This step is subdivided into following sub steps:

(i) **Feature Extraction**

First features are extracted from the pre-processed tokens i.e. tokens are chosen that are considered as feature vector which will go in next phase for classification data is obtained. The common feature extraction methods are bag- of- words, bag -of- n-grams (bigrams, or more), POS (Parts of Speech) Based features.

(ii) **Bag-of-Words**

The feature vector can be represented by a simple approach called bag-of-words in which the text is consider as a group of words stored in a bag in which order of words are not considered.

Thus, the sentence is represented as feature vector

$$d_i = (wi_1, w_{i2}, \ldots, wi_j, \ldots, wi_n)$$

where wi_j is the weight of feature w_i in the sentence d_i, n is number of features in the sentence.

(iii) **Strength of Bag-of-Words**

- It is a simple technique to understand and implement.

(iv) **Weakness of Bag-of-Words**

- The technique represents no spatial relationship between features.
- In BoW, an unordered list of unigrams (single one word) is obtained without any information like related to syntax, semantics etc .So it is less usable, where such informative features are required.

(v) **Bag-of-N-grams**

When feature vector is represented as bag of N-grams, the n contiguous sequence of words in the text from text corpus is together called a n-gram. For n = 1, the n-gram is called "unigram"; for n = 2, the n-gram is called "bigram", for n = 3, the n-gram is

called "trigram", for n> 3, we simply replace the letter n by its numerical value, such as 4-gram, 5-gram, etc. A vector of unigrams is often called the Bag-of-Words model. Consider the sentence "he bought iphone and ipad ". That can be represented as a vector of unigrams [he; bought; iphone; and; ipad]. Besides, this sentence can be represented also be represented as vector of bigrams [[he bought]; [bought iphone]; [iphone and]; [and ipad]].

(vi) **Strength of Bag-of-N-grams**

- The local positioning information can be obtained by extracting n-grams instead of individual words.

(vii) **Weakness of Bag-of-N-grams**

- It is necessary to have a optimum length of n-gram .If length of n-gram is too short, it may not get the important differences, if length is too long, it may get only particular cases and may not generalize well.

(c) POS (Parts of Speech) Based features

When feature vector is represented by POS tagging, it means that each word in the text will be assigned a tag which represents its role in the grammatical context i.e. text is considered as group of verb, noun, adverb or adjective (POS tagged words). The general POS tags are JJ to denote adjectives, RB to denote adverbs, VB to denote verbs, and NN to denote nouns. The combination of adjectives, adverbs, and nouns performed better than individual tokens when considered as feature vector. Adjectives play an important role as indicators of opinions. So they performed the best among the three (i.e. Adjectives, adverbs, and nouns) individual POS tagged features.

(i) **Strength of POS Based features**

- The POS based features can help to deal with the problem of word sense disambiguation. If the correct sense of word/feature is known, it would help in improving the performance of classification.

(ii) **Weakness of POS Based Features**

- When POS tagging is done, many words can have more than one tag, to know the correct meaning according to the context of word/feature can be a tedious job.

(iii) **Feature Weighting**

Apart from using efficient feature extraction method, it is important to assign correct weight to the feature used. In this step weights are assigned to the features according to the importance of sentiment. The common feature weighting schemes are term frequency, binary-weighting scheme, term frequency-inverse document frequency (tf-idf).

(iv) **Term frequency**

It is a feature-weighting technique in which the features are weighted by term frequencies (TF). For example, if a term happy appears in the document 5 times, then the feature term happy is weighted with number 5.

(v) **Strength of Term Frequency**

- It is one of the simplest methods of knowing of feature importance.

(vi) **Weakness of Term Frequency**

- The feature weighting technique does not work when the term features are not just concentrated in only one particular document, but they are frequent in other documents in the corpus. Due to this precision of classification may get affected.

To overcome the above limitation, another feature weighting technique TF-IDF can be used.

(vii) **Binary-weighting scheme**

Another commonly used feature weighting values are binary numbers, which indicate the presence/ absence(1/0) of term in the text corpus, Here the presence/ absence of term is more important than frequency . It means that if a term (feature) appears in the document or sentence, then its weight value in that document or sentence is 1, else is 0. For example, if a term happy appears (no matter how many times) in the document, then the feature weight value of happy is 1, else the value is 0.

(viii) **Strength of Binary -weighting**

- It is a feature weighting technique which can be used in classification where numeric value of term frequency is not used.

(ix) **Weakness of Binary -weighting**

- The binary representation may underperform than other weighting technique where term frequency plays an important role in measuring the importance of feature.

(d) Term Frequency -Inverse Document frequency (TF-IDF)

In this scheme, weights are assigning to each term according to how these terms occur in other document. Rare terms could have high idf score, which means that rare terms may be good indicators for text classifications, depending on how it could balance well with tf scores .Its weights are computed by :

$w_{ij} = tf_{ij} * idf_i$

where tf_{ij} is the term frequency i.e. Frequency of term 'i' in document 'j'

idf_i is the inverse document frequency which is equal to $\log(N/n_i)$,N is the total number of documents in the corpus, and n_i is the number of documents containing the term 'i'.

For e.g. suppose a review contain 100 words wherein the word great appears 6 times. The term frequency (i.e., TF) for great is (6 / 100) = 0.06. Again, suppose there are 1 million reviews in the corpus and the word great appears 1000 times in whole corpus Then, the inverse document frequency (i.e. IDF) is calculated as log(1,000,000 / 1,000) = 3. Thus, the TF-IDF value is calculated as: 0.06 * 3 = 0.18.

(i) Strength of TF-IDF

- The formula required to calculate is easy to understand and implement.
- It gives a different perspective for considering the feature weight age .i.e. it gives importance to not only the terms in the document but in other documents too.

(ii) Weakness of TF-IDF

- As it is a term frequency based weighting measure, it does not provide information about the position of features, semantic-related information.
- Another limitation is that it is not able to recognize the singular and plural form of the word (feature), and consider the singular and plural form of word as two different features.

(iii) Feature Selection

In this step the features extracted in previous sub step are analyzed to select the important features that are not redundant, noise-free and are relevant for classification, since the irrelevant/redundant input feature vector will induce greater computational cost. Feature selection methods are techniques that select important features out of a given set of features using some goodness of a term formula, top features are selected above the threshold criteria and other irrelevant features are dropped. If the length of feature vector is high, the dimension of feature vector is also reduced. The common feature selection methods are Document Frequency (DF), Information Gain (IG), Chi –square (CHI).

(iv) Document Frequency(DF)

This is a simple feature selection method where we find the number of documents in which the feature appears at least once, which is called document frequency of the feature. The terms that are less than a pre-determined threshold are removed.

(v) Strength of DF

- One of the simplest techniques for feature reduction.
- Scalability to a large corpus can be done easily.

(vi) Weakness of DF

- It happens sometimes that low DF feature tends to more informative.

(vii) Information Gain(IG)

Information Gain is frequently employed as a term-goodness criterion in the field of machine learning .The heuristic considered is to select those attribute that is more pure than other attributes in the group. To achieve this entropy of attributes is calculated**,** i.e., of the degree of disorder of the system.

IG(A) is the reduction in the entropy that is achieved by learning a variable A:

$$IG(A) = H(S) - \sum_i \frac{S_i}{S} H(S_i)$$

where H(S) is the entropy of the given dataset and $H(S_i)$ is the entropy of the ith subset generated by partitioning S based on feature A. A feature with high information gain should be ranked higher than other features because it has stronger power in classifying the data.

(viii) Strength Of Information Gain

- As information gain is a good measure for deciding the importance of attribute in the feature vector .It can be used to decide sequence of attributes in the nodes of decision tree.

(ix) Weakness Of Information Gain

- One of the major disadvantages is that information gain is more prone to the attributes with large number of instances in compare to attribute with less number of instances. So accuracy gets affected if the data is unbalanced.
- The above disadvantage can lead to over fitting of data i.e. due to information gain bias; the non-optimal attribute may get selected for prediction.

To overcome this bias, information gain ratio can be used. This method uses the value of split info to normalize the value of information gain. Split info gives information about what proportion of the information gain is actually valuable for that split. The attribute with the greatest information gain ratio is selected.

(e) Chi-Square(CHI)

In this feature selection technique the independence between the two events (class) is examined. The test is used to rank all our terms by their independence with respect to classes and then set a threshold to select only top n features with the highest value of the Chi-square statistic.
It can be calculated as

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

Where Oi is the observed number of cases in category i,
Ei is the expected number of cases in category i.

(i) Strength of Chi Square

- A chi square test can be applied for measuring the 'goodness of fit 'of an actual data that is observed with an expected sample distribution to test the independence of criteria of classification. It gives information that how accurate is the sample data is a representation of actual data.
- The formula required to calculate chi-square test is easy and interpretation of output is also easy to understand.
- The Chi-square statistic is a non-parametric (distribution free) test designed to analyze and test the group differences.

(ii) Weakness of Chi Square

- One weakness or Limitation is with respect to its sample size requirements i.e. a sufficient size of sample is required to get correct approximation from chi-square test.
- Another limitation is the requirement is the data format. The input data must be in the form of frequency table (i.e. data samples with their respective frequencies).

(f) Machine learning based Classification on Feature vector

After the construction of feature vector in the previous phase, the optimal features obtained are classified by applying suitable machine learning classification algorithm. The main classification algorithm considered by researchers in sentiment analysis is Naïve Bayes [11], Decision Tree [12], Maximum Entropy [13], Support Vector Machine [14].

(i) Naïve Bayes

Naïve Bayes classification predicts the probability of a given feature belong to a particular class label by using Bayes theorem. The classification is considered naïve because it is assumed that the features have no dependence on each other i.e. one feature does not affect the other in classifying whether or not the review is positive, negative or neutral.

The Bayes theorem for finding probability, is defined, for a given data point x(i.e .word) and class 'c'(here in case of SA ,c = positive ,negative or neutral):

$$P(c|x) = \frac{P(x \mid c)\, P(c)}{P(x)}$$

P(c|x) =Posterior probability(i.e. the resultant probability of attribute x in test set belong to class label 'c') P(x|c)= Maximum Likelihood or Conditional Probability(i.e. overall probability of attribute x in training set belong to class label 'c')
P(c)=Class Prior Probability
(i.e. probability of c' in training set)
P(x)=Predictor prior probability (i.e. probability of attribute x in all the class labels in training set)

(ii) Strength of Naïve Bayes

- It is simple to implement .i.e. easier to predict class label on test data.
- Training time and prediction time required is less as compared to other text classification algorithms.
- If textual data for training is fairly little, then high bias/low variance classifier i.e. Naïve Bayes classifiers does well in such circumstances. As the interactions between the attributes are ignored in the model, there is no requirement of examples of these interaction and therefore less data is required than other text-classification algorithms.
- The performance of classifier is good with independent feature vectors.

(iii) Weakness of Naïve Bayes

- The performance of algorithm can degrade if the data contains highly correlated features.
- One of the problem encountered in Naïve Bayes is Zero Observation problem .i.e. if a categorical attribute has a value in the test set that was not there in the training set .Then the model will assign a zero probability and be unable to make a prediction .But Some techniques like laplacian smoothing can be applied to overcome zero observation problem.

(g) Decision Tree

Decision Tree is a non-parametric supervised method used for Sentiment Analysis text classification. It classify the given feature to a particular class label by developing a tree model where the given data is divided recursively(according to some condition) until the leaf nodes which contain the nodes for the classified class label. The performance of algorithm depends on choosing the best splitting node which

means choosing the feature which gives maximum information**.**

(i) Strength of Decision Tree

- Simple to understand, visualized and to interpret even by non programmers. Decision trees are white-box classification algorithm means they are able to generate understandable rules in human readable form.
- Simplifies complex interaction between input variable and target output by dividing the original input variables into significant sub-groups.
- Decision tree gives a clear indication of which are the attributes feature i.e. words are most important for classification.
- Decision tree are non-parametric means no specific data distribution is necessary.
- It can easily handle feature interactions and they are robust to outliers.

(ii) Weakness of Decision Tree

- Decision tree learners can create over-complex trees that may not generalizes well this is called over fitting problem. The over fitting decision tree require more space and other computational resource. Pruning helps in handling the over-fitting problem.
- It does not work well with continuous attribute as compared to categorical one.

(iii) Maximum Entropy

Maximum entropy classification algorithm is a probabilistic classifier based on the principle of Maximum Entropy. Here the classifier select the feature goes to a class label that maximizes the entropy. The algorithm does not assume that features are independent with each other.

(iv) Strength of Maximum Entropy

- Perform well with dependent features.
- Ability to combine different kind of statistical dependencies in one uniform framework.

(v) Weakness of Maximum Entropy

- The Performance degrades when the feature vectors applied are independent.
- The Max Entropy requires more time to train comparing to Naive Bayes, primarily due to the optimization problem that needs to be solved in order to estimate the parameters of the model.

(vi) Support Vector Machine

The SVM classifier is a non-probabilistic classifier which classifies the data by mapping the feature vector into a high dimensional vector space and plot a separating line between the class labels. Here the goal of SVM is to find a optimal separating Linear hyper plane which maximizes the margin between the two class labeled data points. Margin is the distance of the closest point of each class from the separating hyper –plane. In order to calculate the margin, two parallel hyper planes are constructed on each side of hyper plane. If the classes are not linearly separable in the high dimensional space, the algorithm will add a new dimension in an attempt to further separate the classes. It will continue the process until it is able to separate the training data into two separate classes using the hyper plane.

(vii) Strength of SVM

- It can deal with documents with high dimensional input space and pick out many of the irrelevant features.
- It has the ability to produce high classification accuracy compared to other text classification algorithm.

Weakness of SVM

- To choose the values of parameters in SVM is hard
- To choose the best kernel function in SVM is also a difficult problem.

(viii) Analysis of Workflow of Sentiment Analysis

The study shows the important phases that need to be considered while performing Machine learning based Sentiment Analysis. The phases such as text pre-processing, feature vector selection affect the performance of any classification algorithm. Different techniques that can be applied with their pros and cons are discussed. The choice of method may vary depending on the application, domain and language used.

In sentiment analysis, in the first phase the decision of how to represent the text is important. For this, it is important to know what type of tokenization to use. If stop word removal technique is to be used, then what stop words need to be considered as list of stop words vary in different domain and language .If stemming is used ,care need to be taken that it may not lead to over or under stemming. For feature vector extraction, if a simplest method is required then bag –of-words is used. A more improved version of bag-of –words is bag- of -n grams, but here length of n should not be too short or too long. POS tagging is also used for feature extraction which provide grammatical context of features.

Feature weighting techniques are also used to rank the features based on their frequency. A simple method known as term frequency is used, when the features are concentrated to a single document, if the terms are distributed in other documents then TF-IDF feature weighting technique is used.

Feature selection method chooses the features that are more relevant than other features. Two commonly used feature selection algorithms in Text Classification are the information gain and the Chi-square test. Each algorithm evaluates the features in a different way and thus leads to different selections. Information Gain is used as goodness of measure by measuring the entropy of the attributes. If the instances of attributes are unbalanced, then the problem of over fitting may be encountered. Chi-Square method is it measures how well the observed distribution of data fits with the distribution that is expected if the variables are independent.

The final step is to classify the data using classification algorithm. Four classification algorithms were discussed .Naive Bayes is a probabilistic classifier based on Bayes Theorem whose performance is good with independent feature vector. The performance of algorithm can degrade if the data contains highly correlated features. Decision tree are simple to understand and can be interpreted by non-programmers. Sometimes decision trees may create over-complex trees that do not generalize well. Maximum entropy classification algorithm work well with dependent features. The Max Entropy may require more time to train comparing to Naive Bayes, primarily due to the optimization problem that needs to be solved in order to estimate the parameters of the model .SVM can deal with documents with high dimensional input space and pick out many of the irrelevant features. To choose the best kernel function in SVM is a difficult problem. As there is no free lunch theorm i.e. there is no single classification algorithm that performs well in all topics/domains/applications. It can't be concluded that SVM are always better than decision trees or vice-versa. The accuracy of a classifier can be as high as 95% in one domain/topic and as low as 40% in some other. It is hoped that by understanding the different phases in the study would help in selection of the classification algorithm, feature selection technique and other configurations to improve the accuracy, precision and recall of algorithm.

IV CONCLUSION

Sentiment Analysis is the process of extracting opinion from textual data. This paper presents a study of workflow of sentiment analysis which would help the researchers in improving the performance of automated Sentiment Analysis system.

REFERENCES

[1] Bing Liu (2010) Sentiment Analysis and Subjectivity, Handbook of Natural Language Processing, Second Edition.

[2] Bo Pang Lillian Lee, and Shivakumar Vaithyanathan, (2002).Thumbs up?: sentiment classification using machine learning techniques. Proceedings of the ACL-02 conference on Empirical methods in natural language processing- Association for Computational Linguistics. Volume 10.

[3] Bo Pang and Lillian Lee,(2008) Opinion Mining and Sentiment Analysis , Foundations and Trends_ in Information Retrieval, Vol. 2, Nos. 1–2 DOI:10.1561/1500000001.

[4] Bing Liu., (2012) Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers.

[5] Vivek Narayanan, Ishan Arora, Arjun Bhatia(2013). Fast and accurate sentiment classification using anenhanced Naive Bayes model. Intelligent Data Engineering and Automated Learning 14th International Conference, Hefei, China.

[6] Rincy Jose, Varghese S Chooralil. (2015). Accurate Sentiment Analysis using Enhanced Machine Learning Models. International Journal of Science and Research (IJSR) Volume 4 Issue 9 ISSN (Online): 2319-7064.

[7] Gurneet Kaur, Abhinash Singla.(2016).Sentimental Analysis of Flipkart reviews using Naïve Bayes and Decision Tree algorithm International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 5 Issue 1, ISSN: 2278 – 1323.

[8] Mohamad Syahrul Mubarok, Adiwijaya, and Muhammad Dwi Aldhi(2017). Aspect-based sentiment analysis to review products using Naïve Bayes International Conference on Mathematics: Pure, Applied and Computation AIP Conf. Proc. 1867, 020060-1–020060-8; doi: 10.1063/1.4994463.

[9] Walaa Medhat, Ahmed Hassan, Hoda Korashy(2014) Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal.

[10] S.Vanaja, K.Ramesh kumar (2014).Analysis of Feature Selection Algorithms on Classification: A Survey. International Journal of Computer Applications (0975 – 8887) Volume 96– No.17.

[11] A.McCallumK. Nigam.(1998)A Comparison of Event Models for Naive Bayes Text Classification , and Learning for Text Categorization.Papers from the AAAI Workshop.

[12] J.R. Quinlan (1986) Induction of Decision Trees.Machine Learning 1,pp-81-106.

[13] Adam L. Berger Stephen A. Della Pietra ,and Vincent J. Della Pietra, (1996).A Maximum Entropy Approach to Natural Language Processing, Association for Computational Linguistics Volume 22, Number 1

[14] Simon Tong and Daphne Koller,(2001) Support Vector Machine Active Learning with applications to Text Classification. Journal of Machine Learning Research, pp-45-66.