Volume 5, Issue 8, August 2015 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Sentiment Classification and Analysis Using Modified K-Means and Naïve Bayes Algorithm

Ashish Shukla[*]
M.tech Scholar, CSE Department
Pranveer Singh Institute of Technology, Kanpur
U.P.T.U., Luck now, Uttar Pradesh, India

Rahul Misra
Assistant Professor, CSE Department
Pranveer Singh Institute of Technology, Kanpur
U.P.T.U., Luck now, Uttar Pradesh, India

Abstract— Sentiments are central to almost all human, actions and activities and can influence our perception and behaviour. People as well as organizations express their sentiments, also called opinions everywhere mostly on internet as the people now days are much dependent on internet. So the requirement of user opinions analysis is gaining importance day by day. People post their experiences, and give feedbacks about the products and services that they are using. Blogs, micro blogs, review sites, twitter, and other social networks are the most common platforms that are used by people and organizations for posting their views. These are rich sources of data that is used in sentiment classification and analysis. Researchers has done very immense effort in the field of sentiment analysis and also new opportunities and challenges still arise so even now it is very active and dynamic research area in the field of natural language processing. It is also widely investigated in text mining, data mining and web mining. We proposed a sentiment analysis system using modified k means and naïve Bayes algorithm that saves running time and reduces computational complexity. The same system can be extended to other product review domains easily.

Keywords— Opinion Mining, Sentiment Analysis, Modified k means, NLP

I. INTRODUCTION

Sentiment analysis is ultimately related to natural language processing. It tracks the public feelings and mood about a certain product or service they are using. People give their feedbacks and share their opinions in blogs, review sites and other social networking sites like Twitter and Face book. Sentiment analysis or opinion mining is used to build a system that collect and analyse feedbacks of customers about the specific product or service. Opinion mining turns to be very useful in many ways. Taking a simple example, in marketing environment let some new product is launched in the market and people are asked for giving their reviews, opinions and experiences after using that particular product. Manufacturer or organization then can analyse the shortcomings of that product, the actual need of the customers and enhance their products accordingly. Sentiment analysis is very important and crucial for market competitors. It helps them in their decision making process. They can identify which particular product or which product feature is more suitable for particular geographic or demographic region. Sentiment classification has many applications in several fields. For example it can be used to classify the product reviews into positive and negative class. This is very helpful for the new customers in gaining the overall idea of what other existing customers are saying about that product so that they can decide whether the product should be bought or not. It can also be used to filter out email messages with abusive and impolite words that can be placed into spam category. One of the major applications of sentiment analysis is Text Categorization. Text classification is the process of classifying written text documents into some categories or classes from a pre-defined training dataset. Text categorization is widely used in many applications related to Natural Language Processing and has gained considerable attention in recent years from researchers as well as the academic and industry developers. Many tools given by Information Retrieval and machine learning systems are being used by Text Classification because it is content based document classification task that shares several properties with information retrieval tasks. There are many opportunities and new challenges are arising continuously in the field of sentiment analysis. There are some basic problems are encountered when we talk about sentiment classification. For example a particular word may have ambiguous appearance that means sometimes it behaves like positive word and sometimes behaves like negative word depending upon the situation. Also traditional text processing process says that small difference among the text documents do not change the overall meaning very much but in sentiment analysis process “the product is good” is far different from “the product is not good”. So it has to be kept in mind that customers express their sentiments in different ways and not always in a same way. Moreover most of the comments or reviews made by people have both positive and negative statements and there may be a contradiction in customer comments. The remaining paper is described in the following sections; section **II** illustrates some well-known sentiment classification techniques, section **III** describes several sources of data i.e. used for sentiment analysis, section **IV** presents the proposed architecture sentiment classification and finally section **V** evaluates derived results and conclusion.

II. SENTIMENT ANALYSIS TECHNIQUES

Generally sentiment analysis can be performed at the following 3 levels: the document level, sentence level, aspect or attribute level [14] [15]. The literature describes two types of techniques called supervised learning and unsupervised learning, in relation to sentiment analysis. Moreover, many techniques and methods of natural language processing is being used here in sentiment analysis more specifically for sentiment classification at the document level. So sentiment detection therefore shares information, knowledge and many properties with information retrieval and natural language processing systems for example text mining, text search predicative analysis, effectiveness measures etc. This section provides brief details of the machine learning and unsupervised learning algorithms used in the experiments.

1. SUPERVISED LEARNING

Since early 2000, researchers have been studying about Machine learning, also known as supervised learning and using this they derived opinions from feedbacks and reviews posted online. Several machine learning techniques have been applied to sentiment classification. The most widely used supervised learning techniques for sentiment classification for product reviews are Naïve Bayes(NB) Classification, Maximum Entropy(MaxEnt), Support Vector Machines(SVM), Neural network, Multi-Layer Perceptron (MLP), Decision tree. This algorithm need training data to perform and for this dataset of labelled opinion words is needed.

Multi-Layer Perceptron (MLP): An MLP is also known as Artificial Neural Network(ANN) .An MLP can be considered as network of neurons called perceptrons. The perceptron computes a single output from multiple inputs.MLP is also known as feed forward networks and can have one or more hidden layers between input and output layer. The MLP networks can be used for both supervised and unsupervised learning process. [2]

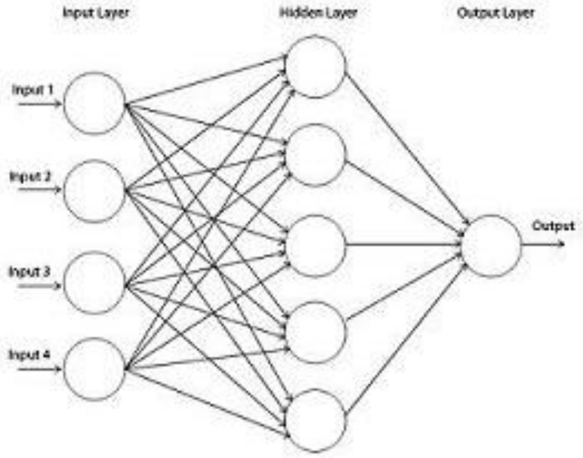


Figure 1:MLP

The above architecture has the following properties: 1.There is no connections within a layer, 2.There is no direct connections from input to output layers, 3.The layers are fully connected, 4.Generally there are more than 3 layers, 5.It not necessary that the no. of input units are equal to the no. of output units, 6.No.of hidden units in each layer can be more or less than input or output units.

The MLP network should have minimum three hidden layers for any valid representation and such a network takes much time for its training process.MLP is the most used type of neural network algorithm and having huge number of applications. It is capable of modelling complex functions. It is very good at ignoring irrelevant inputs and noise and it can be used even with a few knowledge is available about the relationship of the function to be modelled.

2. UNSUPERVISED LEARNING

Unsupervised Learning tries tries to find the hidden structure in unlabeled data. That is why it does not require any prior training in order to analyze the data. Instead of that, it tries to measure how far a particular word is tending towards positive and negative sentiment. This model does not perform well until all the input values are available. If some of input values are missing, it can’t derive anything about the outputs. Several methods have been employed for unsupervised learning in the field of data mining that are used to process the data. Clustering algorithm, expectation-maximization algorithm, matrix factorization, principal component analysis and many others are the common examples. Unsupervised learning can learn models that are having deep hierarchies. It sometimes can be used to cluster the data into categories on the basis of their statistical properties only.

Unsupervised sentiment analysis research and analysis makes use of lingual resources. Kamp’s et al [4] used lexical relationships in sentiment analysis and classification. Andrea Esuli and Fabrizio Sebastiani [5] proposed semi-supervised term classification for determining the orientation of subjective terms. Their basic idea is to do quantitative analysis of the glosses of these terms. When the review have not enough contextual information to determine the actual sentiment, Chunxu Wu[6] proposed a method in which contextual information present in other reviews about the same topic is gathered and analyzed, then by using semantic similarity among them, one can judge the orientation of that sentiment. Ting-Chun Peng and Chia-Chun Shih [7] examined unsupervised learning algorithm. In the proposed work opinion phrases of each document is extracted by applying the rules of part-of-speech patterns. An approach proposed by Gang

Li & Fei Liu [8] is based on the k-means clustering algorithm. This approach used the phenomenon of TF-IDF (term frequency – inverse document frequency) weighting applied on the raw data. After that an efficient clustering algorithm is applied to derive best clustering results. Polanyi and A. Zaenen [10] examined the effect of valence shifters on classifying the sentiments of the documents. Chaovalit and Zhou [9] compared two approaches namely; Semantic Orientation approach and N-gram model machine learning approach .they applied both of these on movie reviews.

III. DATASOURCE

Organizations use sentiment analysis to understand how the public feels about something at a particular moment in time, and also to track how those opinions change over time.. Blogs, micro blogs and review sites serve as rich data sources for sentiment classification and analysis.

Blogs

A blog is a webpage that contains information about someone’s activities or interests. People can read a blog and they can write their own opinion about what it contains. Usually blogs are updated frequently. People exchange their views with one another on the topics they want to discuss on a blog. There are millions of messages are posted at a time and these blogs are used for sentiment analysis. [14]

Micro Blogs

Micro blog is a kind of blog that enables users to broadcast short text messages or media i.e. pictures, video, or sounds to other users of the service. Social networking sites, like Twitter or Face book are the most commonly and widely known examples of micro blogs. Sometimes these Twitter messages express sentiment that can be assumed as the data source for sentiment classification and analysis. [16]

Review Websites

There are plenty of websites are available on the internet in which thousands of consumers are generating reviews for products and services they are using. These reviews play important role in decision making for the new user about what to purchase and what to not. In sentiment analysis and classification customer reviews data is needed that is available on the different websites like www.reviewcentre.com (product reviews), www.fonearena.com (mobile reviews), www.flipkart.com (product reviews), in which thousands of product reviews are available commented by consumers. [15] To figure out the distinguished features, feature selection technique and best supervised learning algorithm, one can use the openly accessible movie review dataset [17]. This classic dataset called as Cornell Movie Review Dataset .It contains two thousand reviews which are having one thousand positive and one thousand negative reviews which are extracted from Internet Movie Database.

IV. PROPOSED METHODOLOGY

The proposed architecture consists of four modules: user interface, pre-processing, Feature Extraction and Clustering using Modified K-means, and Naïve Bayes Classification. Firstly, we use Modified k-means method for feature extraction and clustering. Feature extraction is the practice of choosing a subset of the words appearing in the training database and taking only this subset as the features in text categorization. Feature extraction is used for two main reasons. 1. Size of effective vocabulary is reduced by which we can train and apply a classifier method more efficiently. It is useful for classifiers in which training is expensive, unlike Naïve Bayes. 2. Noise features are eliminated by which feature extraction and clustering increases classification accuracy. Modified k-mean algorithm decreases the complexity and effort of numerical calculations for Naïve bayes algorithm. Secondly, Naïve Bayes theorem is then applied to classify the particular document. This system can handle irrelevant data and increases accuracy by associating Modified K means with Naïve Bayes Classification algorithm.

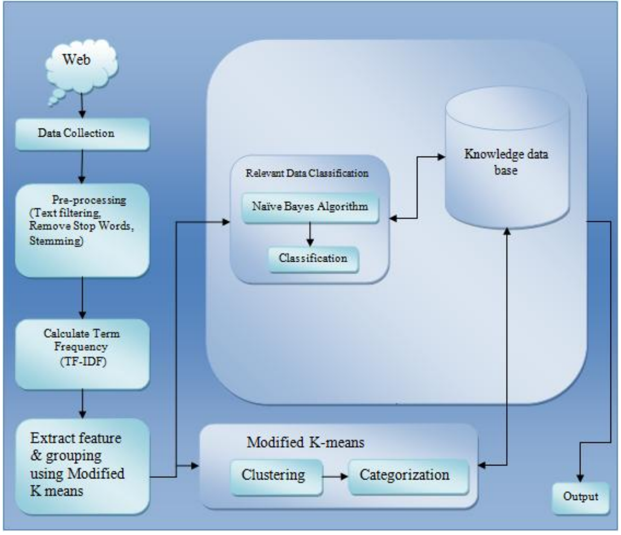

Web
Data Collection
Pre-processing (Text filtering, Remove Stop Words, Stemming)
Calculate Term Frequency (TF-IDF)
Extract feature & grouping using Modified K means
Relevant Data Classification
Naive Bayes Algorithm
Classification
Knowledge data base
Modified K-means
Clustering
Categorization
Output

Figure 2: Proposed Architecture

A. Naive Bayes (NB)

Naive Bayes Classifier, also known as probabilistic classifier is based on Bayes Theorem. It calculates the probability of an instance given the probability of another instance that has already occurred. Mathematically, it can be expressed as;

$$P(C \mid D) = \frac{P(D \mid C)\,P(C)}{P(D)}$$

Where, P (C | D): Probability of Document D being in Class C,
P (D | C): Probability of generating Document D given Class C,
P(C): Probability of occurrence of Class C,
P (D): Probability of document D occurring.

Naive Bayes classifier gives more accurate and efficient results for linearly separable cases and even performs well for non-linearly separable cases [3]. Main advantage of Bayesian Classification is that it can be easily interpreted and it has efficient computation. The algorithm can also be represented using the following figure:

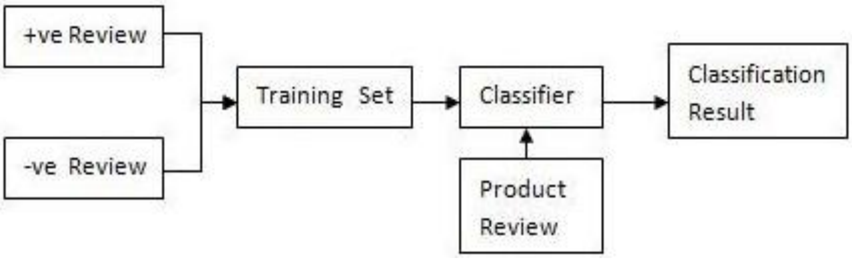


Figure 3: Naïve bayes classification

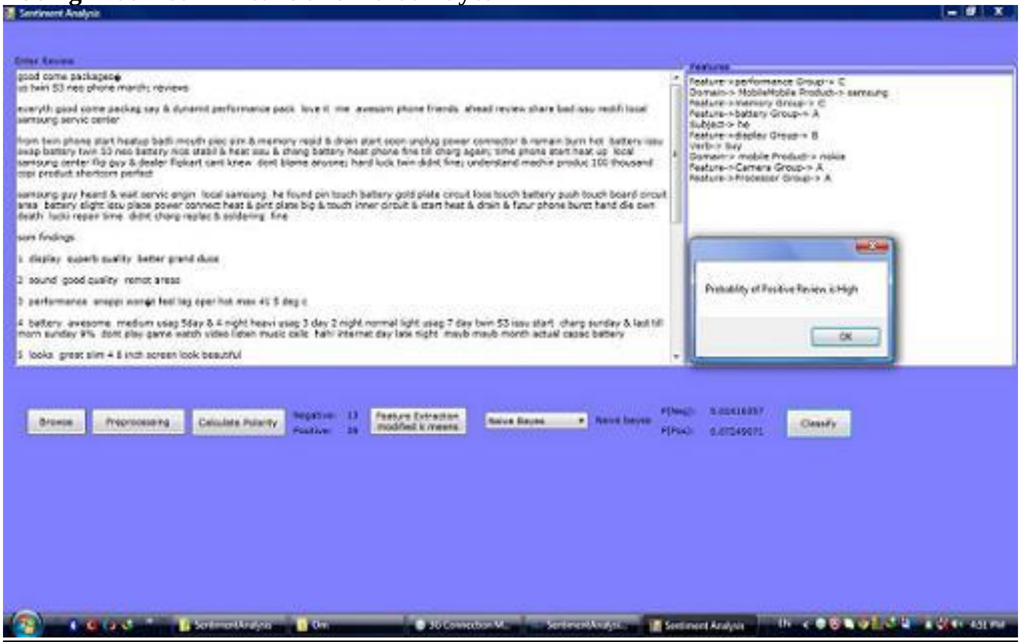
B. Modified approach K-means algorithm:

The K-mean algorithm is widely used for clustering. It is easy to implement and performs well not only for small datasets but can be applied even on large data sets [11]. K-means is a simple algorithm that has been successfully adapted to many applications like computer vision, agriculture, astronomy, market segmentation, image segmentation, bioinformatics, data mining and many others.

For making our experiments scientifically more stable, we are going to use product reviews, more specifically mobile review dataset. So Here we are using product reviews of mobile phones for the experiments We applied above methods on mobile review dataset that contains 2000 various mobile reviews retrieved from Amazon (www.amazon.com), Flipkart (www.flipkart.com), and Review Centre (www.reviewcentre.com) for our experiment. These reviews are available for different types of domains. Each of those domains has 1000 positive and 1000 negative labelled reviews. Out of these 2000 reviews, 1400 reviews are used for training and rest 600 reviews for testing. An average number of words in a particular document are normally greater in Movie review dataset than Product review dataset.

V. RESULT

Classification using Modified K means and Naïve Bayes



good come packagecd
us twin S3 neo phone month reviews

everyth good come packag say & dynamit performance pack love it me awesom phone friends alread review share bad issu redifi local samsung servic center

from twin phone start heatup badli mouth piec sim & memory reaid & drain start soon unplug power connector & remain burn hot battery issu swap battery twin S3 neo battery nice stabil & heat issu & charg battery heat phone fine till charg again time phone start heat up local samsung center flip guy & dealer flipkart cant knew dont blame anyone hard luck twin didnt fine understand machin produc 100 thousand copi product shortcom perfect

samsung guy heard & wait servic engin local samsung he found pin touch battery gold plate circuit loos touch battery push touch board circuit area battery slight issu place power connect heat & pint plate big & touch inner circuit & start heat & drain & futur phone burst hand die own death lucki repair time didnt charg replac & soldering fine

som findings

1 display superb quality better grand duos

2 sound good quality remot areas

3 performance snappi wont feel lag oper hot max 41 5 deg c

4 battery awesome medium usag 5day & 4 night heavi usag 3 day 2 night normal light usag 7 day twin S3 issu start charg sunday & last till morn sunday 9% dont play game watch video listen music calls hahi internet day late night mayb mayb month actual capac battery

5 looks great slim 4 8 inch screen look beautiful

Feature->performance Group-> C
Domain-> MobileMobile Product-> samsung
Feature->memory Group-> C
Feature->battery Group-> A
Subject-> he
Feature->display Group-> B
Verb-> buy
Domain-> mobile Product-> nokia
Feature->Camera Group-> A
Feature->Processor Group-> A

Probability of Positive Review is High

Negative: 13
Positive: 39

Naive bayes
P(Neg): 0.00416357
P(Pos): 0.07249071

Figure 4 :Result Snap

We obtained an overall classification accuracy of 89.01% on the test set. The algorithm takes O (n + V log V) running time to train and O (n) running time to test. We compute accuracy (Manning and Sch¨utze, 1999) of the classifier on the whole evaluation dataset, i.e.:

Accuracy = #No. of Reviews Correctly classified / #No. of Total Reviews Processed

Table1: The characteristics of the evaluation dataset

Name of the Algorithm	**Dataset**	**Accuracy(%)**
Naive Bayes	2000 mobile review dataset	79.66
SVM	2000 mobile review dataset	83.59
Naïve Bayes+Modified K-Means	2000 mobile review dataset	89.01

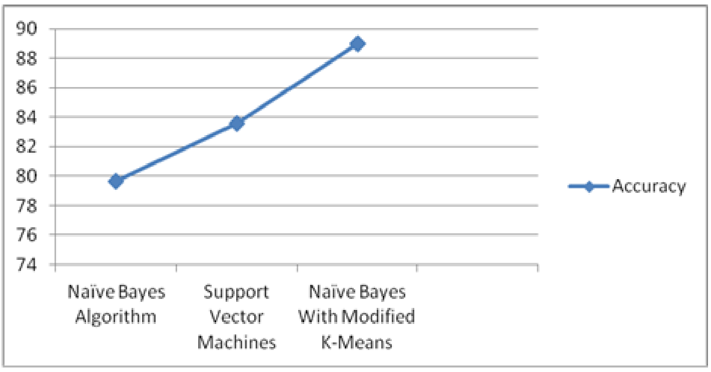


Figure 5: Classification Accuracy

The above graph and table shows the evolution of classification accuracy and how the proposed method helped to increase the accuracy of the classifier.

VI. CONCLUSION

We proposed a method using naïve bayes and modified k means clustering and found that it is more accurate than naïve bayes and support vector machine techniques individually. This study has investigated that proposed method is much quicker than other existing machine learning methods like Support Vector Machines or Maximum entropy which take a much time to give optimal results. The accuracy can be compared to that of the current state of the art algorithms that are used for sentiment classification and analysis on mobile reviews.

From our point of view the combination of MKM and Naïve Bayes gives better results for text based classification and Support Vector Machines for social interpretation. In future we will be focusing to find out how other methods, when applied to customer reviews, can be improved to give more accurate results for sentiment analysis.

REFERENCES

[1] "Sentiment classification using machine learning techniques." Bo Pang, Lillian Lee and Shivakumar Vaithyanathan, In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 79–86.

[2] Towards Enhanced Opinion Classification using NLP Techniques, IJCNLP 2011, pages 101–107, Chiang Mai, Thailand, November 13, 2011

[3] Qiang Ye, Ziqiong Zhang, Rob Law, "Sentiment classification of online reviews to travel destinations by supervised machine learning approaches", Expert Systems with Applications 36 (2009) 6527–6535.

[4] Kamps, Maarten Marx, Robert J. Mokken and Maarten De Rijke, "Using wordnet to measure semantic orientation of adjectives", Proceedings of 4th International Conference on Language Resources and Evaluation, pp. 1115-1118, Lisbon, Portugal, 2004.

[5] Andrea Esuli and Fabrizio Sebastiani, "Determining the semantic orientation of terms through gloss classification", Proceedings of 14th ACM International Conference on Information and Knowledge Management,pp. 617-624, Bremen, Germany, 2005.

[6] Chunxu Wu, Lingfeng Shen, "A New Method of Using Contextual Information to Infer the Semantic Orientations of Context Dependent Opinions", 2009 International Conference on Artificial Intelligence and Computational Intelligence

[7] Ting-Chun Peng and Chia-Chun Shih , "An Unsupervised Snippet-based Sentiment Classification Method for Chinese Unknown Phrases without using Reference Word Pairs", 2010 IEEE/WIC/ACM International Conference on Web Intelligence and intelligent Agent Technology JOURNAL

[8] Hu, and Liu, "Mining and summarizing customer reviews", Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, WA, USA, 2005,pp. 168–177.

[9] Chaovalit, Lina Zhou," Movie Review Mining: A Comparison between Supervised and Unsupervised Classification Approaches", Proceedings of the 38th Hawaii International Conference on System Sciences – 2005

[10] Polanyi and A. Zaenen, "Contextual lexical valence shifters," in Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text, AAAI technical report SS-04-07, 2004.

[11] Shailendra Singh Raghuwanshi, PremNarayan Arya "Comparison of K-means and Modified K-mean algorithms for Large Data-set"

[12] Jin-Cheon Na, Christopher Khoo, Paul Horng Jyh Wu, “Use of negation phrases in automatic sentiment classification of product reviews”, Library Collections, Acquisitions, & Technical Services 29 (2005) 180–191.

[13] Zhongwu Zhai, Bing Liu, Hua Xu and Hua Xu, Clustering Product Features for Opinion Mining, WSDM’11, February 9–12, 2011, Hong Kong, China. Copyright 2011 ACM 978-1-4503-0493- 1/11/02...$10.00

[14] Singh and Vivek Kumar, A clustering and opinion mining approach to socio-political analysis of the blogosphere, Computational Intelligence and Computing Research (ICCIC), 2010 IEEE International Conference.

[15] G.Vinodhini and RM.Chandrasekaran, Sentiment Analysis and Opinion Mining: A Survey, Volume 2, Issue 6, June 2012 ISSN: 2277 128X International Journal of Advanced Research in Computer Science and Software Engineering

[16] Alexander Pak and Patrick Paroubek, Twitter as a Corpus for Sentiment Analysis and Opinion Mining

[17] Pang, B., Lee, and L.: A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of the Association for Computational Linguistics (ACL), pp. 271-278 (2004)