Supervised Machine Learning For Sentiment Analysis

Prashant Kumar Shrivastava

Abstract: Web has been becoming a very important part in people's life. People express their opinions and reviews related to the products and services on the web. Therefore, product reviews are generated daily on large scale. By analyzing these products reviews, new customers find others opinion. The categorization of reviews is very important for any business to grow. Broadly reviews are classified as positive or negative. Sentiment analysis is broadly applied to voice of customer materials like opinions, reviews and responses. Manufacturers or organizations become aware of good and bad things about their products, service and their competitors by analyzing sentiments from reviews of users. In order to make and maintain impression in market, every organization is continuously watching user reviews. In this paper we proposed to classify the sentiments from product reviews using supervised machine learning. Performance of Support Vector Machines (SVM), K Nearest Neighbor (KNN) and Decision Tree algorithms are compared and analyzed.

Keywords: Sentiment Analysis, Supervised Machine Learning, SVM, KNN, Decision Tree

1. INTRODUCTION

As a widespread utilization and growth on the Internet in recent years has created a large quantity of opinionative data available. Especially in the automotive, electronics gadgets, e-commerce, restaurants and movie domain, customers give reviews about products, service or their features. By analyzing these products reviews, new customers find others opinion and experience about various features of the product or service. They can compare the products to each other to find the best one that meet their needs. By analyzing reviews, manufacturers or organization will decide good and bad things about their products or service or those of their competitors. So that, maintain smart customer relations which is required to grow the business. As such opinionated web content is very large, manual analysis of reviews by people and organization for decision making is not possible and time consuming. Taking all these scenarios into consideration there is a need of automating the process of analyzing large data and extracting opinions of customers. The automated process of analyzing opinionated data in form of written language or text, to find out people's opinion is called sentiment analysis which is popular and interesting area for researchers nowadays [10]. Supervised Machine learning approaches such as Support Vector Machine, K Nearest Neighbor and Decision Tree are used to classify data.

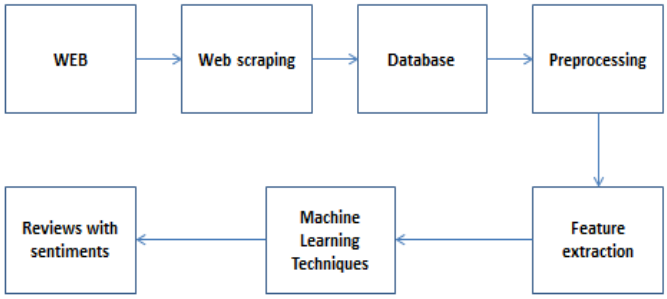
2. LITERATURE REVIEW

In 2009, Daniela XHEMALI et.al compares Neural Networks;Naïve Bayes and Decision Tree classifiers used text classification. In this, Naïve Bayes used to find the aspects in product reviews. It also classifies the polarity of aspects. Naïve Bayes is used with Chi Square technique also POS tagging as a feature selector. The result shows that Naïve Bayes classifier works best for the training courses domain [1].In 2011, F. Pedregosa et.al implemented Scikit-learn explains supervised and unsupervised machine learning algorithms with consistent and task-oriented interface, so enabling comparison of methods for a given application [2]. In 2013, survey conducted by Abhishek Fulmari et.al in which many labeled corpuses used for generating classification and then uses bag-of-words model and three machine learning methods, naïve bayes, maximum entropy classification and SVM for the classification of film reviews. The results depicts that SVM has better performance than other classifiers [3]. In the same year, I. Hemalatha et.al applied Machine learning algorithms: Naive Bayes algorithm, maximum entropy classification for the classification of tweet sentiment with similar performance [4]. In 2014, S. Kiritchenko et.al applied Multi Class Support Vector Machine and dictionary based approach for the classification process. [5]. In the same year, a research conducted by P. Gamallo et.al proved that Naïve Bayes gave better results for sentiment analysis of English Tweets [6]. In the same year, Walaa Medhat et.al conducted survey in which different techniques for sentiment analysis are discussed along with applications and tools for sentiment analysis [7]. In the same year, Ms. Gaurangi Patil et.al develops a classifier for sentiment classification. In this research, people's comments and opinions about political candidates are classified as the positive or negative. It uses (SVM) classifier which gives good performance on text classification [8].In 2015, Xing Fang et.al proposed, a sentiment polarity classification process. Reviews of product are collected from the site Amazon.com. Scikit-learn python package is used. NaïveBayesian, Random Forest and Support Vector Machines classification models are used for categorization [9].In 2016, K. Schouten et.al provides a survey on aspect based sentiment analysis. For this task considers three different types of methods: dictionary based, supervised machine learning, and unsupervised machine learning. In this paper supervised machine learning method is used because supervised data is available and supervised methods work better than unsupervised methods [10].In 2017, Mira Dholariya et.al surveyed various techniques that utilizes within the field of Sentiment analysis, additionally discuss applications and some tools which will be used for sentiment analysis [11]. In the same year, K. Schouten et.al enhances the sentiment analysis using domain ontology information. By incorporating common domain knowledge into ontology, performance of classification for aspect identification and aspect sentiment classification can be improved. The authors found words within sentences that appear in the ontology and are related to the aspect under consideration. They then provided all the super classes of the ontology concept to employed machine learning algorithm for the classification tasks. For both classification tasks works with an existing classifier, the linear SVM. Aspect level sentiment analysis consider at the review-level [12]. Machine Learning Approach:Machine learning (ML) enables systems to analyze or learn from data, examples or experience. After learning stage it makes decisions based upon that data. Machine

learning consists of two types of learning i.e. supervised learning and unsupervised learning. In supervised learning, classification maps data into predefined groups or classes [3]. A supervised learning method, classification algorithm uses training dataset and learns itself from that training data.Test dataset is used to analyze the performance of the algorithm. Varioussupervised machine learning algorithms like Support Vector Machines, K Nearest Neighbor and Decision Tree are used for classification of text data.

3. PROCESS OF SENTIMENT ANALYSIS

1. Data Collection: Customer reviews are periodically crawled from the website using web scrapping technique. These reviews distributed as positive reviews and negative reviews. This collected data is used as training dataset and test dataset for the system.
2. Data Preprocessing: This step is very important. It is the process of preparing and cleaning the data of the dataset for classification. In preprocessing stage stop words, punctuations etc. are removed from the reviews.
3. Feature Extraction: Once the data is pre-processed, features relevant for sentiment analysis are extracted. On the cleaned data, the term weighting module (TFIDF) is applied to form a feature vector space.
4. Training and testing machine learning classifier:After the features extraction, ML algorithms are used for classification of reviews. Figure I depicts process flow for sentiment analysis.



FigureI: *Process Flow for Sentiment Analysis*

4. MACHINE LEARNING ALGORITHM FOR SENTIMENT ANALYSIS

a. K Nearest Neighbor

In KNN, while classifying the new data, the algorithm searches for the nearest neighbors from the training set and use the class labels of the most similar K neighbors to the new data. The KNN stores trained reviews in vector form and the test reviews are classified by using similarity measures on train and test vectors. The KNN algorithm modeled in python. It uses KNeighborsClassifier, provided by SciKit-Learn. For the distance measure between two data points the Euclidean Distance is used.

Euclidean Distance is,

$$dist(X,Y) = \sqrt{\sum_{i=1}^{k} (Xi - Yi)^2}$$

b. Support Vector Machines

In SVM, the group of data is classified using linear separator or hyperplane into different classes. There could be number of hyper planes which classify the data into different classes. Among these hyper planes, the best hyperplane is chosen based on the maximum normal distance between data points. This maximum distance is the margin. During training phase, this algorithm uses hyperplane for separating positive and negative reviews. New reviews are classified based on where on hyperplane it lies. In SVM, maximum margin tends to less misclassification. The SVM algorithm modeled in python using the Python machine learning library scikit-learn.

c. Decision Tree

A decision tree is a popular algorithm for classification. Decision tree is a tree like flow chart, in which each node represents feature, each link or branch represents rule or decision and each leaf represents a category or label. In the starting, the entire training dataset is considered at the root, for selecting the category for the new review data. The root node contains the condition which checks the input reviews features. After that selects a branch which choses that feature value. To model decision tree, gini index split criteria is used. The decision tree algorithm modeled in python. It uses the tree library, provided by SciKit-Learn.

5. PROCESS OF SENTIMENT ANALYSIS

The performance of the system is decided using theaccuracy, precision, recall and F-score. Confusion matrix is used to calculate these measures.

Table I. *Confusion Matrix*

		Predicted	
		neg	pos
Actual	neg	True Negative (TN)	False Positive (FP)
	pos	False Negative (FN)	True Positive (TP)

Accuracy: It is used to measure how often the classifier is correct. Following formula is used to measure accuracy:

Accuracy (%) = (TN + TP) / (TN + FN + FP + TP)

Precision: It is used to measure how accurately the classifier makes predictions w.r.t each class. It is calculated as ratio of true positive examples to the total of false positive and true positive examples.

Precision (%) = TP / (FP + TP)

Recall:It is percentage of positive tuples which the classifier labeled as positive. It is the ratio of true positive examples to the total of true positive and false negative.

Recall (%) = TP / (FN + TP)

F-score: It is calculated as,

$$F - score = \frac{2 * Precision * Recall}{Precision + Recall}$$

Data Set:The data set has been prepared by taking positive and negative reviews by scraping from the consumer affairs site.

6. EXPERIMENTAL RESULTS

After the training of classifier model, the performance of classifier is tested using accuracy, precision, recall and F-score. Accuracy as a metric helps to understand the effectiveness of our algorithm. KNN used with different nearest neighbors (K- value). For different K- values accuracy score varies. 10-NN gives highest accuracy score 78.57%. Accuracy score for different K-values is shown in Table II. Table III depicts the performance of SVM, KNN and Decision Tree. The accuracy is the overall accuracy of algorithms. Dataset includes reviews of two classes. Precision (pos) and Recall (pos) ratios are for positive reviews. Precision (neg) and Recall (neg) ratios are for negative reviews. These results show that SVM classifier achieves higher accuracy (85.71%). KNN classifier gives 78.57% accuracy and Decision tree gives 69.04% accuracy.

Table II. *KNN Results*

Sr. No.	K-Value	Accuracy (%)
1	3	73.80
2	5	76.19
3	7	69.04
4	9	76.19
5	10	78.57

Table III. *Results*

Machine Learning algorithm	precision (%)		Recall (%)		F-score (%)		Accuracy (%)
	neg	pos	neg	pos	neg	pos	
KNN	71.42	92.85	95.23	61.90	81.63	74.85	78.57
SVM	85.71	85.71	85.71	85.71	85.71	85.71	85.71
Decision Tree	75	65.38	57.14	80.95	64.86	72.34	69.04

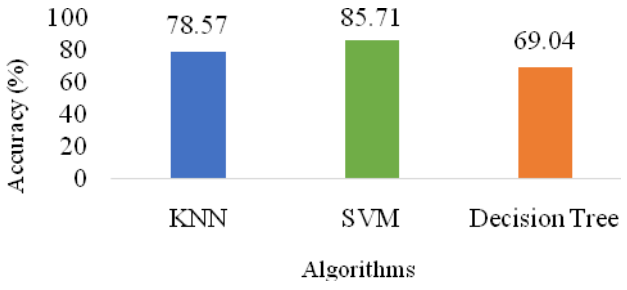


Figure II: *Accuracy of Algorithms*

Figure2 shows the bar graph representation of accuracies of algorithms: SVM, KNN and Decision Tree.

7. CONCLUSION

This paper proactively analyzes supervised machine learning techniques: SVM, KNN and Decision Tree for performing sentiment classification. The performance for sentiment classification is checks out using accuracy, precision as well as recall values and F-score. In this paper performance comparison of machine learning algorithms is done. The experimental results show that for given dataset the K Nearest Neighbor giving 78.57% accuracy for k=10. Decision tree gives 69.04% accuracy. It is observed that SVM performs better than KNN and Decision Tree with 85.71% accuracy. Future scope of this work consists of implementation of ensemble classifier instead of supervised algorithm and comparing performance parameters accordingly. Also use different ways for analyzing performance such as cross validation, bootstrap, etc. In Future multiclass classifier could be used to determine different sentiments like positive, negative, or neutral

REFERENCES

[1] D. Xhemali, C. J. Hinde and R. G. Stone, “Naïve Bayes vs. Decision Trees vs. Neural Networks in the Classification of Training Web Pages” in International Journal of Computer Science Issues, vol. 4, no. 1, pp. 1623, 2009.

[2] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python” in Journal of Machine Learning Research, 12, pp. 2825– 2830, October 2011.

[3] Abhishek Fulmari and Manoj Chandak, “A Survey on the Supervised Learning for Word Sense Disambiguation” in International Journal of Advanced Research in Computer and Communication Engg, Volume 2, Issue 12, Dec 2013.

[4] I.Hemalatha, G. P Saradhi Varma and A.Govardhan, “Sentiment Analysis Tool using Machine Learning Algorithms” in International Journal of Emerging Trends &Technology in Computer Science (IJETTCS), Volume 2, Issue 2, March – April 2013.

[5] S. Kiritchenko, X. Zhu, C. Cherry, and S. Mohammad, “NRC-Canada-2014: Detecting aspects and sentiment in customer reviews,” in Proceedings of the 8th International Workshop on Semantic Evaluation (Association for Computational Linguistics and Dublin City University, Dublin, Ireland), pp. 437–442, 2014.

[6] P. Gamallo and M. Garcia, “Citius: A Naïve Bayes Strategy for Sentiment Analysis on English Tweets” in Proceedings of the 8th International Workshop on Semantic Evaluation (Association for Computational Linguistics and Dublin City University, Dublin, Ireland), pp. 171–175, 2014.

[7] Walaa Medhat and Ahmed Hassan, “Sentiment analysis algorithms and applications: A survey” in Shams Engineering Journal, volume 5, pp. 1093–1113, 2014.

[8] Gaurangi Patil and Kalpana Dange, “Sentiment Analysis Using Support Vector Machine” in International Journal of Innovative Research in Computer and Communication Engineering, Volume 2, Issue 1, January 2014 .

[9] Xing Fang and Justin Zhan, “Sentiment analysis using product review data” in Journal of Big Data, pp. 2:5, 2015.

[10] K. Schouten and F. Frasincar, “Survey on Aspect Level Sentiment Analysis” in IEEE Transactions on Knowledge and Data Engineering, Volume 28(3), pp. 813–830, 2016.

[11] Mira Dholariya, Dr.Amit Ganatra, Prof. Dhaval Bhoi, “A Survey on Sentiment Analysis: Tools and Techniques” in International Journal of Innovative Research in Computer and Communication Engineering, Vol. 5, Issue 3, March 2017.

[12] K. Schoutenand F. de Jong, “Ontology-Enhanced Aspect-Based Sentiment Analysis” in Proceedings of the 17th International Conference on Web Engineering (ICWE 2017), Volume 10360, pp. 302–3320, Springer, 2017.