Analysis of investigation regarding sentiment analysis applying machine learning approaches

Veera Swamy Pittala[1], Ramesh Babu Pittala[2]

Assistant Professor[1], Professor,HOD,CSE[2]
[1]Lakireddy Bali Reddy College of Engineering Mylavaram
veeraswamypittala@gmail.com
[2]KLR College of Engineering and Technology Paloncha
prameshbabu526@gmail.com

***Abstract*—** To evaluate whether a text includes subjective information and what information it communicates (i.e. whether the attitude underlying the text is good, negative, or neutral), is the job of Natural Language Processing known as Sentimental Analysis. Several machine learning methods for sentiment analysis and opinion mining are the subject of this study. Predicting product reviews and customer attitude toward a newly introduced product may be possible with the use of sentiment analysis combined with machine learning. This article provides a comprehensive overview of the available machine learning methods, which are then contrasted in terms of their accuracy, benefits, and drawbacks. When compared to unsupervised learning methods, the 85% accuracy we get with supervised machine learning is a significant improvement.

Keywords*— Emotional analysis, Discriminators, Learned data, Data-free learning, SVM, and Supervised Learning;*

INTRODUCTION

Using sentimental analysis, one may ascertain how individuals feel about their own unique life. Opinion, emotion, sentiment, attitude, opinions, and conduct expressed by users in online reviews have become commonplace. Sentiment analysis is a technique for determining the emotional tone of written material. Sentimental analysis is an approach of analyze the body language and underlying ideas of a keynote speaker, author, or author vs an aggressive field or object. Sentiment analysis makes several assertions. The first is a point of view that will be considered unfavorable in another situation when it was considered good. The second argument is that individuals seldom think about their perspective from the same angle. You may piece together the review's mixed bag of praise and criticism by analyzing each statement on its own. It may be time-consuming to track down and identify online opinion sites. Therefore, a summary method and robotic opinion mining will be required.

Classification at the document level, the phrase level, and the feature level are the three tiers of sentiment analysis. The primary goal of document-level categorization is to determine if the overall tone of a document is favorable or negative. It makes assumptions about the whole text at once. Sentence-level analysis aims to classify the feelings conveyed by individual sentences. The first stage in identifying whether a statement is objective or subjective occurs at the sentence level. If a statement is subjective, the person reading it gets to determine whether the viewpoint expressed is favorable or bad. Aiming to classify feelings towards specific things, this approach is used in aspect-level analysis.

In the field of sentiment analysis, two main methods predominate. Both symbolic and machine learning approaches are taken into account. Learning techniques in symbolic learning include those based on analogy, discovery, examples, and root learning, among others. Unsupervised learning, semi-supervised learning, and supervised learning are all used in machine learning. Machine learning will be viewed as one of the key approaches in sentiment categorization, alongside lexicon-based and linguistic methods. Sentiment analysis and categorization methods are shown in Fig.1.

1.1 Machine Learning Approach

Machine learning is a branch of AI that uses algorithms to teach computers how to learn on their own. Unsupervised learning, semi-supervised learning, and supervised learning are all used in machine learning.

1.1.1 Instructional Guidance

Learning a function from an experiment's input and output is part of a supervised machine learning approach that is also associated with the usage of a labeled feature set to preserve some categorization. In supervised learning, you are given a dataset that has been previously tagged with the target function. Each training example in the training data set comprises of a pair of input data and a predicted result.

Learning with Limited or No Supervision 1.1.2.

Since labeled corpora are required for these supervised approaches to work, they are not always applicable in practice. Weakly-supervised and unsupervised approaches to machine learning, which do not rely on manually labeled data, are another possibility. In weakly supervised learning, the amount of labeled data is relatively low in comparison to the amount of unlabeled data. When using an unsupervised technique, the input is used as a learning device.

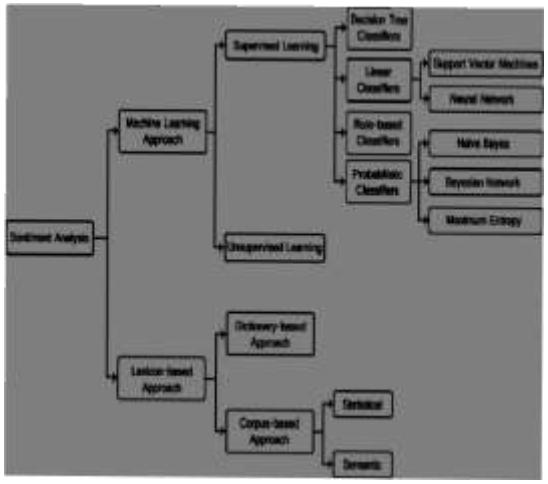

Figure 1 Sentiment classification techniques.

and no estimates of the projected yield are provided. Cluster analysis and expectation-maximization algorithms are two types of unsupervised learning methods. These computations gather emotional text through a Dictionary-based method. Every word in a dictionary has another term that is the antonym of it. down this way, we may leverage the dictionary's antonym and synonym arrangements to zero down on a suitable emotional seed word. At first, only a handful of words are gathered, and their positive or negative coordination is established. This process is repeated until no further words can be discovered.

1.2 Methodology Utilizing a Lexicon

Calculating and labeling the weights and counts of words related with sentiment in a lexicon is what makes the lexicon-based approach to sentiment classification possible. Dictionary-based approaches, corpus-based methods, and the manual opinion approach are all taken into account while compiling the perspective list.

I. Sentiment Classification Based on MachineLearning Methods

In machine learning technique it uses unsupervised learning, weakly supervised learning and supervised learning.

Classification using a Decision Tree 2.1

Features were assigned to internal nodes in the Decision Tree classifier, and the outgoing edges were labeled as trials based on the data set weight. The names of the tree's leaves are organized hierarchically. This classification scheme classifies the whole text from top to bottom, branch after branch, until it reaches a leaf node. Decision tree learning uses a decision tree classifier as a predictive model, mapping data about an item to inferences about that item's likely quality. A big quantity of input can be figured out in a limited period of time employing reliable computing resources in decision trees. The key benefits of using a decision tree classifier are that it is straightforward and simple to comprehend. Minimal data preprocessing is required for this classifier. But these ideas may result in convoluted trees that are hard to generalize.

A Linear Predictive Model

Linear classifiers employ linear decision margins to categorize input vectors into classes. The family of linear classifiers is rather large. As an example, there is the support vector machine. The linear scatters across classes are rather excellent in this classifier.

Neuronal 2.2.1 Network

This neuron is the fundamental building block of a neural network. Non-linear margins were employed in a multi-layer neural network. The output of one neuron is fed into the input of the next neuron in the stack. Because errors must be back-propagated across several layers, training a data set for this form of classifier is more involved.

2.2.2 Stochastic Neural Network

When it comes to voice classification challenges, Support Vector Machine (SVM) is often regarded as the gold standard classifier. To do this, they constructed a hyperplane where the closest trained instances were the furthest apart Euclidean. A small fraction of the training data sets are used to fully resolve the hyperplane in a Support Vector Machine. The certified classifier is not available to the remaining training data sets. As a result, support vector machines (SVMs) have been effectively implemented for text categorization, and they have also been employed in many sequence processing applications. Since SVMs may be utilized without a labeled training data set, they find application in hypertext and text categorization.

Classification Rules 2.3

Rule based classifiers, as the name suggests, combine the creation of a data collection with the development of a set of rules. The left half of the rule specifies the requirement of the aspect set, while the right side specifies the name of the class.

Predictive Classifier 2.4

Many different approaches to classification are used by these classifiers. This plethora of shapes includes all possible types. The probabilities of checking different words for each element are provided by the productive model, which includes all possible elements. Generative classifiers are another name for these systems. Naive Bayes, Bayesian Network, and Maximum Entropy are all examples of probabilistic classifiers.

Naive Bayes Classifier, Version 2.4.1

One common supervised learning approach that serves as a go-to for classification tasks is the Naive Bayesian classifier. Their classifier is considered naïve since it assumes that interconnected events have no effect on one another. The sum of the feasibility reports for each individual word in the document is the meat of the overall document calculation. Since these Nave Bayesian classifiers need less processing power than other methods, they have found widespread use in the classification of emotional reactions. However, their reliance on independence assumptions may lead to misleading conclusions.

A Bayesian Network

The fundamental drawback of the Naive Bayes classifier is that it makes no assumptions about the order in which features appear in data sets. Bayesian networks were first implemented on the basis of this notion. The nodes of this Bayesian network represent variables, while the edges represent the conditional independence between them. Bayesian Network is seldom utilized in text categorization due to its high computational cost.

Superior Entropy

The Maximum Entropy classifier makes use of a weight set to encode a training data set's associations with the joint-future. Because it works by directly extracting certain data sets from the input binding them, the Maximum Entropy classifier belongs to the same family as other classifiers as the log-linear and exponential classifiers.

Classifier with 2.5 K-Nearest Neighbors

To classify texts, K-Nearest Neighbor is an unsupervised learning technique. This method use a number of training data sets to classify entities based on their closest distance to those entities. The simplicity of this method for text classification is one of its main benefits. Multi-class text categorization is another area where this method excels. The primary problem of KNN is that it requires a great deal of time for classifying entities in situations when a big data collection is present.

In table 1 it shows the comparative observation between different machine learning techniques.

II. Literature Survey

From the past set of years, many articles, papers andbooks have been written on sentimental analysis. At the sametime some researchers focus more on specific burden like finding the subjectivity expression, subjectivity clues, subjective sentence, topics, and sentiments of words and extracting sources of opinions, while others target is on

#	Machine Learning Classifier	Advantage	Disadvantage
1	KNN	It is simple and also used for multiclass categorization of document.	It requires more time to categorize when huge number data are inclined. Takes lot of memory for running a process
2	Decision Tree	This is very fast in learning data set. Easy for understanding purpose	It has problem that it is difficult handle data with noisy data Over fitting of data
3	Naïve Bayesian	Simple and work well with textual as well as numerical data. Easy to implement Computationally cheap	Performs very poorly when feature set is highly correlated. It gives relatively low classification performance for large data set.

			Independent assumption of attribute may lead to inaccurate result.
4	Support Vector Machine	High accuracy even with large data set Works well with many number of dimensions No over fitting	Problems in representing document into numerical vector

Table 1: Comparison between machine learning methods In paper [6], [7], and [8] authors proposed aspect based

Emotional research. In the publication [6], the SVM model was put to the test using four different data sets. Both the recall and accuracy rates of the SVM technique and the Maximum Entropy classifier method for feature extraction have been analyzed and found to be better by the authors. The authors of article [7] present a new method that integrates Senti-WordNet with dependency parsing and document-wide sentiment analysis. All sentiment analyzers have included a number of techniques for making instantaneous predictions about the tone of a document or a set of words. Online data from sources like movie reviews, product ratings, and social media are all fair game for sentiment analysis. Some of the methods they use include pattern recognition, NLP, and machine learning. Resolutions including co-references are neatly categorised for emotional analysis. The classifier Support Vector Machine was used to do this. Naive Bayes, Max Entropy, Boosted trees, and Random Forest are some of the most popular sentiment analysis methods, and this research [8] compares and contrasts them. It seems that improved accuracy and performance may be attained when utilizing Random Forest Classifier for sentiment categorization. In addition, this classifier is intuitive and would become better over time in terms of performance. The accuracy was increased at a faster pace due to the aggregation of decision trees. The classifier's considerable processing capacity and lengthy training period are necessary for this purpose. The authors of the research come to the conclusion that the Random Forest classifier, despite its lengthy learning curve, should be used most of the time. The Naive Bayesian classifier was used since it requires less resources and less memory than other methods. In contrast, the Max Entropy classifier is adopted despite its enormous memory and processing time requirements and short training time. These studies show that when it comes to classifying product evaluations, support vector machines perform better than other methods. Sarcastic and comparative sentences, however, have not been addressed in the literature.

In the papers [9], [10], and [19], the authors combine the use of machine learning with NLP to examine a movie review. The authors of study [9] analyzed audience reactions to movies using a support vector machine and a Naive Bayes classifier. Based on this classification, they determine that the linear Support Vector Machine is more accurate than the Naive Bayesian. In their study [10], the authors show how machine learning was utilized to decipher the reviewer's Malayalam. Support Vector Machine and Conditional Random Field (CRF) are used in conjunction with a rule-based method to sentiment classification. In

In [19], the author evaluates the two most popular supervised machine learning methods for review sentiment classification: support vector machine (SVM) and naive bayes (NB). The findings reveal that the support vector machine (SVM) misclassified more data points than the Naive Bayes strategy, and that the Naive Bayes approach beat the SVM with less reviews. The authors concluded that there is a lot of room for development in the areas of corpus construction, efficient feature preprocessing, and feature selection. The automatic analysis of movie reviews' ratings and scores is still a work in progress.

In their papers [11], [12], [13], and [14], the authors detail the numerous methods and programs used to analyze Twitter data for sentiment. It is vital to assess the feelings in Twitter posts since user feedback might be favorable, negative, or neutral depending on the context. Although the authors of study [11] employed lexicon-based approaches for categorization, this approach takes little work for each tagged text content. The study [12] provides an overview of the best practices and the most current developments in the same area. The authors draw the conclusion that supervised learning is superior than unsupervised machine learning when it comes to sentiment categorization. Several sentiment analysis instruments and methods for text categorization are outlined in article [13]. They combine lexicon-based and machine learning methods into one in this hybrid methodology. The combined methods provide for improved categorization accuracy. One of machine learning's primary applications is in adapting and delivering certified design and content for certain purposes. The authors of the study [14] recommended using a Naive Bayesian classifier to evaluate the texts. Their experiments reveal that the Naive Bayesian classifier model performs well on a large data set consisting of lengthy comments and on a variety of social networking sites.

Problems arise when attempting to automate the analysis of tweets because (i) the same term may be interpreted as subjective in one context and as objective in another. (ii) same punishment, different offence (iii) sarcastic utterances (iv) cases when just a fragment of text conveys the whole argument, hence the entire utterance is disregarded (v) Unlike never, never, not, etc., a negative word may be expressed in a variety of ways. It's difficult to make sense of a paradox like that. This implies that these obstacles can be used to make Twitter analysis even better.

Existing methods for assessing sentiments of unstructured data posted on social media were explored in papers [15], [1], [5], and [16]. When analyzing emotions, it ignores neutral language. The authors suggest a method for categorizing texts based on their sentences. SVM, Naive Bayes, Part of Speech, and SentiWordNet were used for this purpose in [15][1]. They get the conclusion that Naive Bayesian and Support Vector Machine classifiers, both examples of machine learning, are the most effective. And perform the role of universally applicable categorization standard. However, the lexical approach has a highly combative tone. A deep learning method was developed specifically for this issue. Classifiers like support vector machines (SVM) and naive bayes (NB) outperform a lexicon-based approach in terms of accuracy in this comparison.

Support Vector Machine (SVM) classifiers are used on benchmark feature sets to evaluate the sentiment classifier, as shown in the papers [5], [16]. N-grams and various weighting schemes were used to extract the traditional characteristics of the dataset. Chi-Square weight features are used to pick features for use in categorization when asked for. The current method's framework consists of preprocessing, aspect selection, aspect extraction, and data set classification. Good results are achieved in text categorization because SVM has the capacity to store large data sets. Another benefit is that SVM works well with little data. The SVM classifier takes as input n-gram, unigram, and other weighting methods. Using these weighting strategies, training the classifier on several common data sets is routine. According to the results of the experiments, the unigram model is superior to the bigram and n-gram models. The authors recommend utilizing the Chi-Square aspect selection strategy to enhance classification accuracy.

The author of article [17] examines the state of the art in sentiment analysis by comparing lexicon-based and machine learning approaches.

along with a strategy that works across languages and domains. Both the lexicon-based strategy and the machine learning approach are shown to be highly competitive and need substantial human effort in document labeling.

Table 2 shows machine learning approach SVM yields highest accuracy as compared to Naïve Bayes and Senti-WordNet.

	TP	FP	FN	TN	Accuracy
Senti-WordNet	148	91	52	109	64.25%
NB	156	81	44	119	68.75%
SVM	135	51	65	149	71.00%

Table 2: Performance comparison of learning methodswith Senti-WordNet

Below table 3 shows performance contrast between different sentiment classification approaches.

	Method	Data set	Accuracy
Machine learning	SVM	Movie reviews	86.40%
	CoTraining SVM	Twitter	82.52%
	Deep learning	Standard benchmark	80.70%
Lexicon based	Corpus	Product Reviews	74.00%
	Dictionary	Amazon	---
Cross-lingual	Ensemble	Amazon	81.00%
	Co-Train	Amazon, IT168	81.30%
	EWGA	IMDb moview review	>90%
	CLMM	MPQA, NTCIR,ISI	83.02%
Cross-domain	Active learning	Book, DVD, Electronics, Kitchen	80% of average
	Thesaurus		
	SFA		

Table 3: Performance comparison of sentimentclassification technique

In paper [18], they have taken online movie reviews for analyzing sentiments. For classification they used three supervised learning approaches such as Naïve Bayes, SVM and kNN. Experimental results show that SVM method beat the kNN and Naïve Bayes approaches. Table 4 shows the collected reviews for sentiment classification.

# experiment	Positive	Negative	Total
1	50	50	100
2	100	100	200
3	150	150	300
4	200	200	400
5	400	400	800
6	550	550	1100
7	650	650	1300
8	800	800	1600
9	900	900	1800
10	1000	1000	2000

Table 4: Collected reviews

The accuracy obtained by using three algorithms are shown in table 5. They have done 10 experiments for each approach. Result shows that even data is either small or large SVM provides higher accuracy than NB and kNN.

# experiment	# reviews	SVM (%)	Naïve Bayes (%)	kNN (%)
1	50	60.07	56.03	64.02
2	100	61.53	55.01	53.97
3	150	67.00	56.00	58.00
4	200	70.50	61.27	57.77
5	400	77.50	65.63	62.12
6	550	77.73	67.82	62.36
7	650	79.93	64.86	65.46
8	800	81.71	68.80	65.44
9	900	81.61	71.33	67.44
10	1000	81.45	75.55	68.70

Table 5: Accuracy obtained after testing data set

III. Conclusion

An overview of recent research on categorizing and analyzing emotions is provided in this study. The results of this study indicate that supervised learning techniques, such as Naive Bayesian and Support Vector Machine, are widely used. The Support Vector Machine outperforms several other classifiers when it comes to accuracy. We found that Naive Bayes works well with a small feature set, whereas SVM is the best option if a big feature set is used. Because they need extensive human intervention with the source documents, lexical techniques are often more forceful. The superior performance of Maximum Entropy is matched only by its vulnerability to over fitting. Despite the fact that many studies have adopted opinion mining using a variety of methodologies, there is still a need for automated analysis that tackles all of the difficulties of emotional analysis at once. In order to overcome obstacles like the categorization of indirect viewpoints, comparison phrases, and sarcastic statements, more creative and effective strategies need to be developed.

REFERENCES

[1] "Machine Learning and Lexicon-based Methods for Sentiment Classification: A Survey", 978-1-4799-5727-9/14 $31.00, Hailong Zhang, Wenyan Gan, and Bo Jiang. © 2014 IEEE.
The following is a summary of "Sentiment analysis algorithms and applications: A survey" by Walaa Medhat a,*, Ahmed Hassan b, and Hoda Korashy, which was published in the Ain Shams Engineering Journal in 2014, volume 5, pages 1093-1113.

[2] Xing Fang* and Justin Zhan, "Sentiment analysis using product review data", Fang and Zhan Journal of Big Data (2015) 2:5 DOI 10.1186/s40537-015-0015-2

[3]"A Survey of Internet Public Opinion Mining" (Kaijie Guo, Liang Shi*, Weilong Ye, Xiang Li), ISBN: 978-1-4799-2030-3 (Hardcover), $14 (Paperback), $31.00 ©2014 IEEE.

[4] "Sentiment Analysis Using Support Vector Machine" by Nurulhuda Zainuddin and Ali Selamat (978-1-4799-4555-9/14/$31.002014 IEEE).

[5]This is based on the work of Chuanming Yu, who presented the paper "Mining Product Features from Free-Text Customer Reviews: An SVM-based Approach" at iCISE 2009, held from December 26-28, 2009, in Nanjing, China.

[6] Raisa Varghese and Jayasree M., "Aspect Based Sentiment Analysis Using Support Vector Machine Classifier", 978-1-4673-6217-7/13/$31.00_c 2013 IEEE, accessed on April 19, 2015.
The following is a citation for "Comparative Study of Classification Algorithms used in Sentiment Analysis" by Amit Gupte, Sourabh Joshi, Pratik Gadgul, and Akshay Kadam in the International Journal of Computer Science and Information Technologies, Volume 5 Issue 5 (October 2014), pages 6261-6264.

[7] "Feature Selection And Classification Approach For Sentiment Analysis", Machine Learning and Applications: An International Journal (MLAIJ) Volume 2, Issue 2, June 2015, authors Gautami Tripathi1 and Naganna S.
[10] 978-1-4799-8792-4/15/$31.00_c-2015-IEEE Deepu S. Nair, Jisha P. Jayan, Rajeev R.R, Elizabeth Sherly, "Sentiment Analysis of Malayalam Film Review Using Machine Learning Techniques,"

Published in the April 2016 issue of the International Journal of Computer Applications (0975-8887), Volume 139, Issue 11 is "Sentiment Analysis of Twitter Data: A Survey of Techniques" by Vishal A. Kharde and S.S. Sonawane.
[8] "A Survey on Feature Level Sentiment Analysis" by Neha S. Joshi and Suhasini A. Itkat in the International Journal of Computer Science and Information Technologies, Volume 5 Issue 4 (April 2014), pages 5422–5425.
The article "Approaches, Tools and Applications for Sentiment Analysis Implementation" by Alessia D'Andrea, Fernando Ferri, Patrizia Grifoni, and Tiziana Guzzo was published in the September 2015 issue of the International Journal of Computer Applications (0975-8887).
[9] Shun Yoshida, Jun Kitazono, Seiichi Ozawa, Takahiro Sugawara, Tatsuya Haga, and Shogo Nakamura, "Sentiment Analysis for Various SNS Media Using Naive Bayes Classifier and Its Application to Flaming Detection",978-1-4799-4540-5/14/$31.00 ©2014 IEEE .
The following is a research paper citing the work of Jalaj S. Modha*, Prof. & Head Gayatri S. Pandi, and Sandip J. Modha: "Automatic Sentiment Analysis for Unstructured Data", IJARCSSE Volume 3, Issue 12, December 2013.

[10] "Sentiment in twitter events," by M. Thelwall, K. Buckley, and G. Paltoglou, published in J. Amer. Soc. Inf. Sci. Technol., vol. 62, no. 2, pages 406-418, February 2011.
For more information on this topic, I recommend reading "Machine Learning and Lexicon based Methods for Sentiment Classification: A Survey" by Hailong Zhang, Wenyan Gan, and Bo Jiang (978- 1-4799-5727-9/14; $31.00). 10.1109/WISA.2014.55 (2014, IEEE).
Sentiment Classification of Movie Reviews Using Supervised Machine Learning Approaches, by P. Kalaivani and Dr. K.L. Shunmuganathan, ISSN: 0976-5166, Volume 4, Issue 4, August/September 2013.

Based on the article "Sentiment Classification Using Machine Learning Techniques" by Suchita V. Wawre1, Sachin N. Deshmukh2 published in the April 2016 issue of the International Journal of Science and Research (IJSR), Volume 5 Issue 4.

[11] "Sentiment analy- sis algorithms and applications: A survey" by Walaa Medhat a,*, Ahmed Hassan b, and Hoda Korashy in the Ain Shams Engineering Journal, volume 5, issue 11, pages 1093-1113.

According to "A Comparative Study on Different Types of Approaches to Text Categorization" by Pratiksha Y. Pawar and S. H. Gawande in the International Journal of Machine Learning and Computing, Volume 2, Issue 4 (August 2012), this is possible.

[12] S. B. Kotsiantis, "Supervised Machine Learning: A Review of Classifi- cation Techniques," Informatica 31 (2007) 249-268 249.