Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 8324 - 8332
Received 05 March 2021; Accepted 01 April 2021.

Twitter Sentimental Analysis Using Augmented Naive-BayesAlgorithm

C.Viji[1], N.Rajkumar[2], R.Sivakumar[3], N.Karthikeyan[4]

[1]Associate Professor, Department of Electronics and Communication Engineering, Akshaya College of Engineering and Technology, Coimbatore.

[2]Associate Professor, Department of Computer Science and Engineering, Akshaya College of Engineering and Technology, Coimbatore.

[3]Professor, Department of Mechatronics Engineering, Akshaya College of Engineering and Technology, Coimbatore.

[4]Associate Professor, Department of Computer Science and Engineering, Sri Krishna College of Engineering and Technology, Anna University, Coimbatore.

Abstract

Sentimental analysis is used in text mining. Twitter is one of the prominentsocial media. Twitter offers organizations a quick and active way to analyze customerviewpoints toward the critical to success within the marketplace. Natural Language Processing, algorithms like the Support vector machine, Naive Bayes is employed to predict the polarity of a sentence. Sentiment scrutinity of Twitter data may be classified upon sentence and document level. The outcomes classify customer perspective via tweets into positive negative and neutral comments, which is represented in a pie chart. This method mostly used in the Market Analysis to the prediction about a product or review about a product. In our proposed method the Naive Bayes algorithm is used for effectiveness and faster processing.

Keywords: Sentimental analysis,Machine Learning,Augmented Naïve Bayes Algorithm, Natural language processing.

1. Introduction

Sentiment analysis is the method to analyze the various customer’s emotions which may be good, bad or neutral from the text data employing text analysis methods. Sentiment analysis tools provides businesses and enterprises to identify the feedback of customers towards brands, products or services based on the feedback acquired trough online. Sentiment analysis is an approach used to scrutinize the text sequence and categorize it into varying labels based on the users input from the social media. The result will be in the form of positive or negative response. In business perspective, sentiment analysis is used to evaluate the effect of their product or ad promotion or response of end users towards their recent updates on social media. By this only they make ananalysis or judge about a product and they will make changes to the product using the comments. This kind of product review helps in market reach and improves the sales of the product.It is impossible to review each and every review. By entering the hash tag, itself we will get a whole analysis how much percentile it is positive and negative.

KDT or Knowledge-Discovery in Text (KDT), refers to the methodology of attaining non-trivial information and knowledge through extracting and stimulating from unstructured script. Text mining can be proposed as an interdisciplinary field which deals with retrieval of data and its processing along with machine learning techniques, linguistics and statistics.Almost 80 to 90% of the information is stocked as text which facilitates text mining to uphold its high commercial significance. Knowledge can e attained from various data sources. Yet, unstructured texts continue to be the most vitalwillinglyaccessible source of data. The dilemma of data Discovery from Text is to mine direct and indirect semantic correlations between concepts using Natural language processing (NLP) method. Iturge insights into heftysize of text data. KDT considered as deeply plunged in NLP, upholds on the techniques from machine learning, statistics, information extraction, reasoning, management of knowledge etc for its breakthrough process. This makes it a noteworthy role in the trending applications including text acknowledging.

1.1 Machine Learning (ML)

Machine learning considered as an efficient technique to train the system and find the optimum solution. Proposed as an appliance of artificial intelligence (AI), itgrants systems the capability to learn and get improvised from previous experience. It is anlinguistics and automated process.It focuses on programs which can access data and employ it to learn for themselves and produces optimum solutions.

Learning process initiates with data (observations), which includes examples, direct experience or instructions so as to observe patterns based on data and to take efficient decision in the future in view of the of the provided examples. The prime intend is to permit computers to automatically get skilled without any participation or support of humans and accordingly amend the actions.

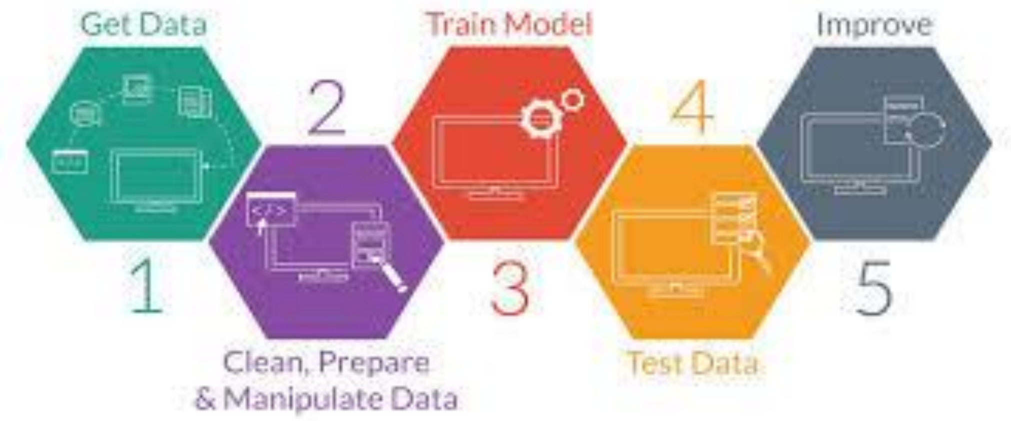


Figure 1.1 Steps in Machine Learning

1.2 Machine Learning Methods

The machine learning offers an autonomous approach for the investigation of multimodal and high dimensional data by preparing refined and automatic algorithms. Machine Learning can be sorted into various types including Deep, Evolutionary, Supervised, Semi-Supervised, Unsupervised and Reinforcement learning.

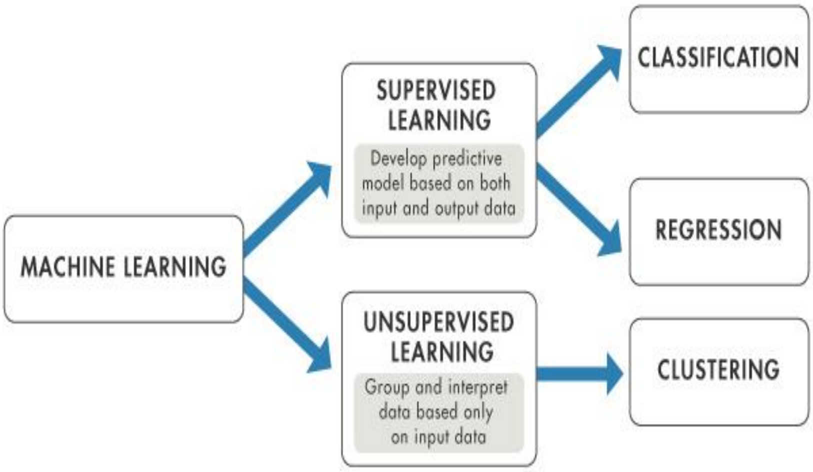

MACHINE LEARNING
SUPERVISED LEARNING
Develop predictive model based on both input and output data
UNSUPERVISED LEARNING
Group and interpret data based only on input data
CLASSIFICATION
REGRESSION
CLUSTERING

Figure 1.2 Machine learning algorithms

Supervised Machine Learning

Supervised machine learning techniques employs labeled examples to apply what model have learned from past to new data and predict future behavior. This algorithm creates an indirect function to make predictions about output based on scrutiny of a known training dataset. After enough preparation, the system will provide target value for any data. This also correlated its attained results with the correct output vales and enhances the model performance accordingly.

Unsupervised Machine Learning

Dataset used for training Unsupervised machine learning technique is neither classified nor labeled and it this methodology learn how a model can recognize a function to illustrate hidden structure from unlabeled data. It is impossible to produce exactoutcomes;however it determines data and can illustrateimplications from datasets. Thisportrays hidden configuration from unlabeled records.

Semi-Supervised Machine Learning

Semi-supervised machine learning algorithms is a combination ofunsupervised and supervised learning. This method applies unlabeled and labeled training datasets. This algorithm exploits a minimal labeled and maxima unlabeled data. It also enhances the accuracy of system learning. Generally, this algorithm is preferred when labeled data necessitates expert and allied resources so as to train or learn from it. Or else, obtaining unlabeled data usually doesn’t entailbonus resources.

Reinforcement machine learning

Reinforcement machine learning technique works by generating and realizing errors based on its environment. Trial and error search and delayed reward are the most important features of reinforcement learning. This method allows machines and software agents to automatically find the ideal behavior within a specific context in order to improve the performance. Simple reward feedback is required for the agent to learn which action is best, this is known as the reinforcement signal.

Evolutionary Learning

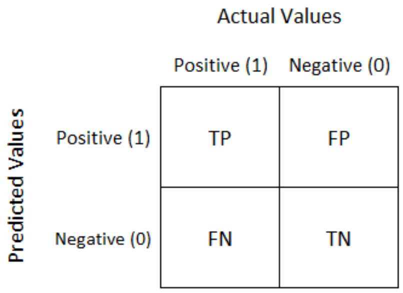
Evolutionary Learningcan be related to a biological system wherein repetitive study occur similar to biological entities so that they can improve their survival rates and gets reproduced. This model works based on idea of fitness based on which its accuracy for a system is evaluated.

Deep Learning

Deep Learning technique is based on algorithm sets and this technique portrays high level data abstraction. Deep graph is employed y this techniques which includes varying processing layers based on which linear an non linear transformation occurs.

1.3 Confusion matrix in machine learning

A confusion matrix can be proposed as a table used to illustrate a classification model's output on a collection of test data for which its true values are known. It enables the revelation of an algorithm's output.



		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Figure: 1.3 Confusion matrix

True Positive(TP):forecasted positive and it’s true.

True Negative(TN): forecasted negative and it’s true.

False Positive (FP): forecasted positive and it’s false. It is a Type 1 Error.

False Negative (FN): forecasted negative and it’s false. It is a Type 2 Error

2. Literature Survey

Revathi S and Rajkumar N, proposed about opinion mining which is employed to analyze and evaluate public opinion for product review. The author classified the sentimental analysis with many algorithms and the performance of accuracy is measured for all the algorithms. In this technique some of the drawbacks are identified but still Low level sentences are not performed for the sentimental polarity.

Revathi S, Rajkumar N and Sathish S explains with Hadoop network. Here the author uses the Rule based algorithm and the ku’s algorithm to evaluate the result. Rule based is analyzed based on the web data. Here the author uses big data, Hadoop, R-BSA, sentimental analysis. In that author says that future enhancement to introduce the new category as neutral.

G. Kavitha and B. Saveen, NomaanImtiaz explains that the skilled persons are only required for acquiring live streaming dataso as to carry out run analysis and henceforth generate the reports. In this study, apache flume and hive warehouse has been employed along with Hadoop wherein the database based on twitter data applications is generated with the aid of Hive and the attained data is imported to a Hive Table. This can be exploited to carry out the real time analyzing of public opinion including governmental actions, elections and various other society based interests.

J. Kampsand et al addresses the issues by concentrating on the mining and reviewing of adjective evaluation classes with the aid of certain adjective evaluation code such as “boring” or “beautiful” and the same can be altered bin accord to the requirement with the aid of certain modifiers which includes “very” or “sort of” or “not”. In owe to the expressions based on Martin and Whites Appraisal Theory, certain taxonomies where adopted for this study and the same was generated based on Systemic Functional Linguistics tradition.Semi-automatic techniques was exploited to develop a lexicon wherein 1329 adjectives and modifiers was gathered and classified to classes of various appraisal attribute taxonomies. From texts, adjectives based appraisal groups were extracted heuristically and their attribute values were computed in accord to the lexicon. Vectors of comparative frequency characteristics were computed from the groups and the discriminating behavior was classified with the aid of a support vector machine learning techniques.

3. Problem

The problem in sentimental analysis is classifying popularity of specified text and document.There are many problems in getting a review about the products or aboutthecompany’s. Consider all the reviews which are posted by the customers. Decisions are depending upon the reviews only based on thearticulatedoutlook of a document being positive, negative or neutral.

4. Existing Algorithms

4.1 Naïve Bayes Algorithm

It can be proposed as a statistical scheme for classification capable enough to resolve problems concerning categorical and continuous value attribute.

$$\mathrm{P(c/d)} = \frac{\mathrm{P(c)\ P(c/d)}}{P(d)}$$

‘d’ is assigned as the initial text classification. Theclass, c* = argmax$_c$ P(c|d)

Here, P(c) has no specific role in opting the c*. fi’sisirrelevant after a’s gets decomposed and the resultant class can be

$$P_{NB}(c|d) = \frac{P(c)\left(\prod_{i=1}^{m} P((f_i|c)^{n_i(d)}\right)}{P(d)}$$

4.2 Support Vector Machine

Adopting a Kernel based algorithm, Support Vector Machine technique is exploited to identify the text similarity and this works mainly based on the training dataset.For example words ‘good’ and ‘excellent’ upholds the same positive polarity. These words can be used to identify the resemblance among them using

$$\vec{w} = \sum_j \propto_j c_j \vec{d}_{j,} \quad \propto_j \geq 0$$

Where,$\vec{w}$ is a hyper plane vector.

αj's is attained through decipheringdual optimization problem.dj so that αj is foreversuperior than zero is said as support vector.

4.3 Rule Based Sentimental analysis algorithm

New algorithms have been proposed is called as Rule based sentimental analysis algorithm. This uses the rule mining algorithm, in this algorithm all the inputs must be provided by the human. About 90% accuracy will be attained through this method wherein opinion is automatically extracted with the aid of lexical dictionary.

5. Comparison with Other Techniques

A review on diverse technique assessable and its variation amid them is as given below.

S. No	Machine Learning Techniques			
	Etiquette	Objectives	Achievements	Restrictions
1	NB SVM MAXENT	Attain validation of three cross fold	Best Performance in accuracy	Unreliable for low frequency characteristics
2	ULT PMI-IR	Verify adjectives semantic orientation	Domain	For large dataset, rate words based distant search engine
3	LCT	Mining Outlook	Automatically classify product features	Strength of opinions and accuracy rate and reliable

Table 5.1 comparison between various algorithms

6. Existing System

Consider Learning Based approach initially, the major positive is that expert knowledge is not mandatory to develop related bases wherein based on the context, it is simply trained.Comparing text sizes is not significant as even though the text size varies, the feature vector sparseness changes for a properly trained classifier. In this technique, texts will be classified base on sizes, and then text and sentence level training will be carried out individually.A large set of training data with both positive and negative level examples are requisites of this method as proposed by Turney which makes it an expensive and time consuming methodology. Along with this, learning and evaluation of classification standards at various clause levels has to be carried out. With amendments in users and publishing time,

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 8324 - 8332
Received 05 March 2021; Accepted 01 April 2021.

the expression also tends to get varied making the training data set to provide insufficient characteristics of the whole data set. This leads to complexity in identifying and correcting the training data set representative.

7. Proposed Work

In this work, through developinga specific system, feedback about a product is being given in a perfect percentage ratio. So we can understand the feedback about a product and can decide whether it is positive, neutral or negative. Here we are performing this system by naives Bayes algorithm

7.1 Algorithm

Naives Bayes classifiers often known as classification algorithm set which is developed based on Bayes Theorem. Naïve-Bayes, a simplest technique of sorting done through Bayesian network wherein all the attributes are independent based on its variable category. An approach to enhance Naive-Bayes is through increasing the structure so as to clearly represent attribute dependencies.

Augmented Naive-Bayes (ANB) can be portrayed as a family of algorithms with common principle and is an extended Naive-Bayes technique. In this method, category node points directly to all or any attribute nodes generating a link between them. This Algorithm provides the accurate result in analysis the sentiment.When compared to other algorithms the Naives Bayes algorithm produces the accurate results. The proposed algorithm is implemented in python.

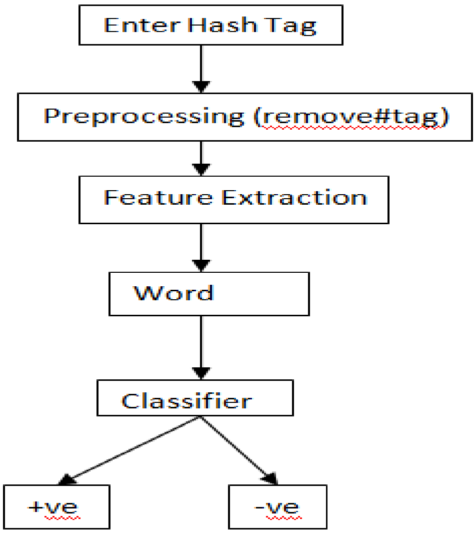


Figure 7.1 ANB Flow Diagram

8. Experimental Results

The Output will be in the form of the pie chart. Whereas the total chart is of 100% and it splits into different regions based on the analysis of tweets;positive, neutral and negative.

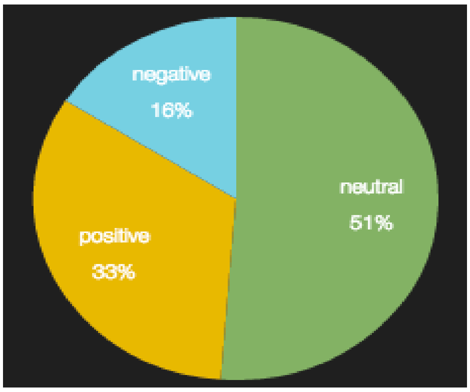


Figure 8.1 Experimental Result

Parameters considered to evaluate the performance are mainly precision, accuracy and recall. And the result will be produced in the form of pie chart so that it is easy for the people to understand the overall review of the product. Comparisons of results achieved from experiments were done and proposed work was established to be superior to present system.

9. Conclusion

This research focused mainly on sentimental analysis, adjectives semantic orientation and accuracy rate performance based on which certain hitches were documented. Results proved that enriching the test's stylistics characteristics and accuracy rate performance has to be done through exploiting Naives Bayes algorithm. To make use of the bigger data set to improve accuracy, of the emotions and sentiments. This study also proposes a future scope wherein feature extraction of new sets more sensitive for early detection has to be done. It also proposes to enhance the detection techniques performance through reducing inappropriate and irrelevant characteristics from existing systems.

References

1. A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining," in Proceedings of LREC, vol. 2010, 2010.
2. E. Boiy, P. Hens, K. Deschacht, and M.-F. Moens, "Automatic sentiment analysis in on-line text," in Proceedings of the 11th International Conference on Electronic Publishing, pp. 349-360, 2007.
3. G. Vinodhini and R. Chandrasekaran, "Sentiment analysis and opinion mining: A survey," International Journal, vol. 2, no. 6, 2012.
4. Z. Niu, Z. Yin, and X. Kong, "Sentiment classification for microblog by machine learning," in Computational and Information Sciences (ICCIS), 2012 Fourth International Conference on, pp. 286-289, IEEE, 2012.
5. L. Barbosa and J. Feng, "Robust sentiment detection on twitter from biased and noisy data," in Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pp. 36-44, Association for Computational Linguistics, 2010.

6. A. Celikyilmaz, D. Hakkani-Tur, and J. Feng, "Probabilistic model-based sentiment analysis of twitter messages," in Spoken Language Technology Workshop (SLT), 2010 IEEE, pp. 79-84, IEEE, 2010.
7. Y. Wu and F. Ren, "Learning sentimental influence in twitter," in Future Computer Sciences and Application (ICFCSA), 2011 International Conference on, pp. 119-122, IEEE, 2011.
8. Revathi S, Rajkumar N, Sathish S, “R-BSA Algorithm of Text Sentiment Analysis for Web-Based Traffic Data” International Journal of Applied Engineering Research, ISSN 0973-4562 Vol. 10 No.1 (2015) pp. 1164-1168, 2015.
9. Revathi S, Rajkumar N, “A Survey on Sentiment Analysis for Web-Based Data by UsingVarious Machine Learning Techniques” International Journal of Engineering Trends and Applications, ISSN 0973-4562 Vol. 1 No.2 (2014) pp. 1-5, 2014.
10. Aveksa Inc. Ensuring “Big Data” Security with Identity and Access Management. Waltham, MA: Aveksa, 2013
11. C. Whitelaw, N. Garg, and S. Argamon, “Using appraisal groups for sentiment analysis,” in Proc. 14th ACM Int. Conf.Inf. Knowl. Manage., 2005, pp. 625–631.
12. J. Kamps, M. Marx, R. J. Mokken, and M. De Rijke, “Using WordNet to measure semantic orientations of adjectives,” inProc. Int. Conf. Lang. Resourc. Eval., 2004, pp. 1115–1118.
13. Jianping Cao, Ke Zeng, Hui Wang, Member, IEEE, JiajunCheng, FengcaiQiao, Ding Wen, Senior Member, IEEE, andYanqing Gao, Member, IEEE” Web-Based Traffic Sentiment Analysis: Methods and Applications”, April 2014.
14. L.-W. Ku, Y.-T. Liang, and H.-H. Chen, “Opinion extraction, summarization and tracking in news and blog corpora,” inProc. AAAI Spring Symp.,Comput. Approaches Anal. Weblogs,2006, pp. 100–107.
15. Q. Ye, W. Shi, and Y. Li, “Sentiment classification for movie reviews in Chinese by improved semantic oriented approach, “in Proc. 39th Annu.HICSS, 2006, pp. 1–5.
16. S.-M. Kim and E. Hovy, “Determining the sentiment of opinions,” in Proc. 20th Int. Conf. Comput. Linguist., 2004,pp. 1367–1373.