Natural Language Processing for Classifying Text Using Naïve Bayes Model

Shailja Joshi

Masters Scholar, Computer Science and Engineering, Geetanjali Institute of Technical Studies, Rajasthan, India.

Mayank Patel

Head of Department, Computer Science and Engineering, Geetanjali Institute of Technical Studies, Rajasthan, India.

***Abstract*: Sentiment examination is the method of programmed identification of the conviction or the mind-set of a creator towards a specific subject in literary structure. To extricate the supposition present in text, the machine needs skill in the territory of characteristic language handling. In this paper, AI based record level slant grouping is performed on Amazon item surveys to characterize them as positive and negative. Two NLP based component extraction procedures (Word Relation and POS based) are utilized in this investigation to decide the highlights that are conclusion bearing. The highlights are extricated as fundamental highlights (unigrams, bigrams and trigrams) and their mixes in request to recognize the highlights that are generally useful and to cut down the computational time of the order calculations, include choice procedures are utilized. Execution of free and joined capabilities is surveyed utilizing exactness, accuracy, review and F-measure. From the investigations led, it is seen that joined highlights outflanked autonomous highlights utilizing Naive Bayes (NB) classifier.**

Keywords: Naïve Bayes, Natural language, Support vector machine, Sentiment analysis

I. INTRODUCTION

The procedure of extraction of valuable data and examples from huge measure of put away information is known as information mining. There are different names for this procedure also, for example, information disclosure process in databases (KDD), data preparing, information extraction or information/design examination. This procedure is otherwise called information digging, information fishing, and information sneaking around. Different kinds of information are examined with the assistance of specific information mining instruments. The huge measure of information which needs certain ground-breaking information examination apparatuses are in this way put for the here which is otherwise called the information rich however data helpless condition [1]. There is an expansion in the development of information, its social event just as putting away it in tremendous databases. It is no more in the possession of people to do it effectively or without the assistance of examination instruments. There are sure information documents made here which can be visited when the information is required [2]. The astute, fascinating and novel examples of information are found from enormous scope informational collections utilizing the information mining. The information disclosure in databases process is a significant advance in information mining. Feeling mining can be described as a sub-control of computational phonetics that focuses on separating individuals' assessment from the web [3]. The current extension of the web urges customers to contribute and convey by methods for sites, chronicles, and relational collaboration destinations, etc. All of these stages give a huge measure of important data that scientists are intrigued to separate. Customer's supposition is an essential standard for the improvement of the nature of administrations rendered and upgrade of the expectations. Sites, audit locales, data and scaled down scale web journals give a fair understanding of the gathering level of the items and administrations [4].

Assessment mining closes whether client's perspectives are sure, negative, or nonpartisan about a specific item, theme, occasion, etc. Assessment mining and rundown process incorporate three essential advances, first is Opinion Retrieval, Opinion Classification and Opinion Summarization. Survey Text is recuperated from systems administration locales.

Set	Document	Review Sentence	Class
Training Set	1	I liked the movie	pos
	2	It's a good movie. Nice story.	pos
	3	Hero's acting is bad but heroine looks good. Overall nice movie.	pos
	4	Nice songs. But sadly boring ending.	neg
Test Set		I like the direction. But boring locations. Overall good movie	pos

Figure 1: Basic Review class

Assessment text in diaries, examinations, comments, etc. contains emotional report about theme. Comments named they are negative or positive audits. Assessment rundown is created taking into account highlights conclusion sentences by thinking about incessant highlights regarding a subject. Different classifiers are used inside the sentiment mining process.

1.1 *Naïve Bayes Classifier:* The most well-known methodology in the hypothesis of administered parametric classifiers is the quadratic segregate work which uses the Bayesian methodology [5]. The target here it to propose a standard which permits appointing the future articles to a class when a lot of items is given for each class.

1.2 *Support Vector Machine (SVM) Classifier:* It is the primary goal of SVM to decide the best capacity by boosting the edge between the two classes. This is because of the way that there are numerous such straight hyper-planes. The measure of room or separation among two classes is known as hyper-plane [6]. The most limited between the closest information focuses to a point on the hyper-plane is known as edge.

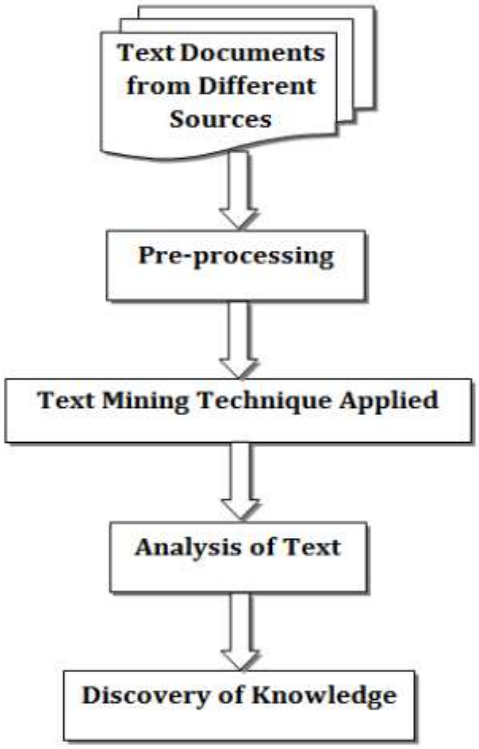


Figure 2: Text Mining Flow

1.3 *Decision Tree Classifier: Decision Trees (DTs):* are a non-parametric managed learning technique used for arrangement and relapse. The objective is to make a model that predicts the estimation of an objective variable by taking in straightforward choice guidelines accumulated from the information highlights. It is a technique for approximating discrete-esteemed objective capacities, in which the educated limit is addressed by a decision tree.

1.4 *K-Nearest neighbor:* These classifiers rely upon learning by relationship. The preparation tests are delineated by n dimensional numeric traits. Each example speaks to a point in a n-dimensional space [7]. Thusly, most of the preparation tests is put away in a n-dimensional example space.

1.5 *Multi-layer Perception (MLP):* The for the most part used feed forward ANN is the multi-layer recognition classifier. To give basic calculations, at first a solitary concealed layer is used. The quantities of neurons are fundamentally included for the improvement of the procedure.

II. LITERATURE REVIEW

LI Bing, et.al, (2014) expressed in this paper [8], lion's share of these works couldn't precisely extricate away from of overall population assessment from the uncertain online life information. They moreover came up short on the ability to sum up multi-qualities from the dispersed mass of social information and use it to gather helpful models. This paper proposes a novel network-based calculation to decide the characterized multilayered Twitter information. They moreover did not have the ability to sum up multi-attributes from the dissipated mass of social information and use it to aggregate valuable models, in like manner did not have any productive component for dealing with the Vast data. Dhanalakshmi V., et.al, (2016) investigated inside this paper [9] feeling mining using directed learning calculations to find the extremity of the understudy input dependent on pre-characterized highlights of educating and learning. The investigation led includes the use of a mix of AI and regular language preparing procedures on understudy input information accumulated from module assessment study aftereffects of Middle East College, Oman.

Every feeling instrument is not quite the same as other. The decision of wistful device to be utilized is totally rely upon client and his/her need. This classifier will have the option to decide the tweets as positive, negative or nonpartisan. This encourages the end client to outline a definitive supposition on the question search. Pooja Kherwa, et.al, (2014) proposed a methodology is this paper [12] that determinedly examines each line of information individually, and produce a pertinent rundown of each audit (ordered by perspectives)

close by different graphical decisions. A one of a kind uses of this technique is assisting item makers or the legislature in gaging reaction. The paper intends to improve our framework as talked about in the earlier segment, and further all the more getting ready many pilot tests to additionally upgrade the outline aftereffects of the framework to create in more detail.

III. RESEARCH METHODOLOGY

This work depends on the conclusion mining in which the highlights of the info information are characterized utilizing the SVM classifier.

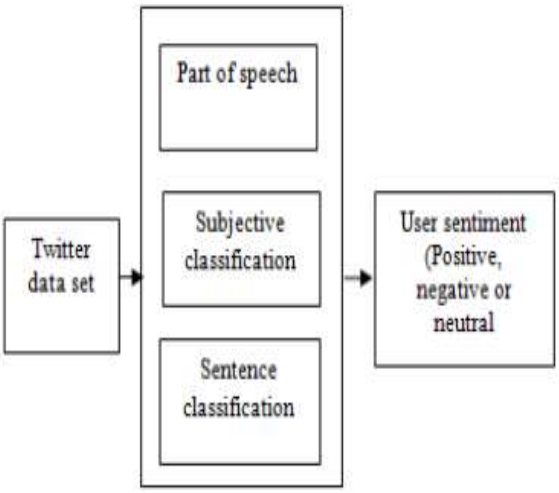
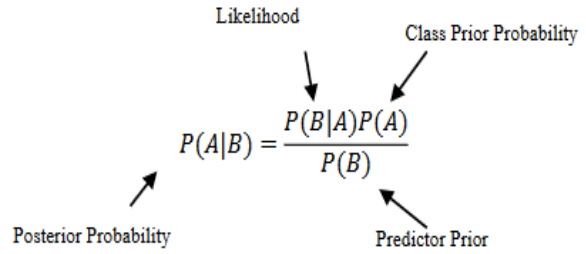


Figure 3: Sentimental Analysis Process

The SVM classifier can be supplanted with the credulous Bayes classifier which is less mind boggling and generally proficient when contrasted with SVM classifier. A basic strategy which is utilized for building the classifiers is known as the Naïve Bayes classifier procedure. For certain difficult cases, the models are made which dole out class names to those issues. Regardless of whether they have a credulous structure and misrepresented suspicions the outcomes are proficient. Just few preparing information is required for evaluating the boundaries which are required for grouping. This is a significant value of the Naïve Bayes procedure.

We utilized a Naive Bayes classifier to characterize the tweets. Guileless Bayes depends on the suspicion of restrictive autonomy among the highlights, some falsehood here. While Naïve Bayes classifiers figure out how to perform well regardless of this presumption, a classifier not dependent on this suspicion may beat a Naive Bayes classifier (Gamallo and Garcia, 2014). The Naive Bayes classifier utilized Laplace smoothing.



$$P(A|B)=P(B_1|A)\times P(B_2|A)\times P(B_3|A)\times \ldots\ldots\ldots\ldots P(B_n|A)\times P(A)$$

Figure 4: Naïve Bayes Equation

A. The framework stream is appeared in the figure underneath. In this framework there is an insightful system to locate the fitting information. Gathering of tweets/post/information from different online networking destinations, for example, twitter. This relates to the stage 1 In Figure 5.

B. Next in the stage 2, we perform Preprocessing on the unstructured tweets. Preprocessing is the way toward cleaning the information from undesirable components. It builds the exactness of the outcomes by lessening mistakes in the information. There are many general preprocessing methods, of which the most widely recognized are: tokenization, secretive content to lower or capitalized, expel accentuation, evacuate numbers, evacuate rehashed letters, and evacuate stop words, stemming and invalidation.

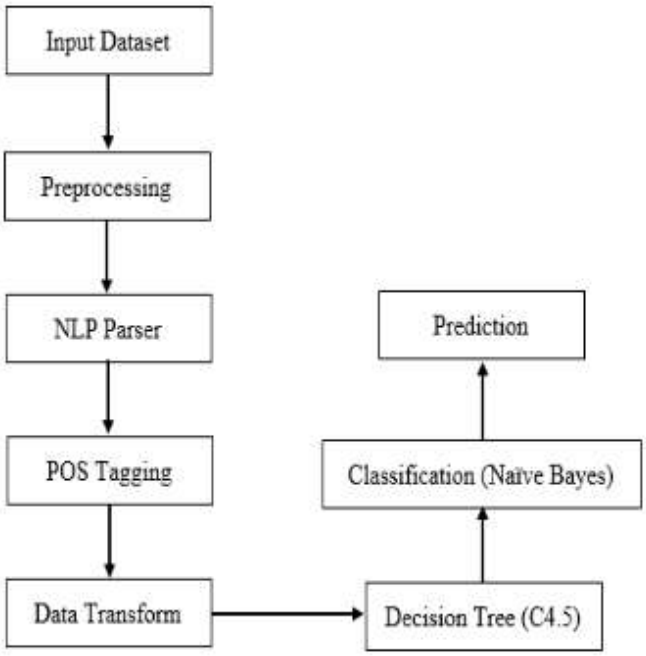


Figure 5: Proposed Method

C. After in the stage 3, a characteristic language parser principally program that works out the syntactic structure of sentences, for example which gatherings of words go together as expressions, subjects, item or action words.

Part of discourse: It is the procedure increasing content as indicated by grammatical form. It utilizes thing, action word, intensifier, descriptive words for distinguishing proof of words.

D. In following stage, we change the information for perception. In information change we for the most part convert information one structure into other. Normal models incorporate sifting and gathering of information.

E. After that we use classifier C4.5 Decision trees are a successful technique for administered learning. Its points is the parcel of a dataset into bunches as homogeneous as conceivable as far as the variable to be anticipated. It takes as information a lot of grouped information, and yields a tree that takes after to a direction graph where each end hub (leaf) is a choice (a class) and each non-last hub (interior) speaks to a test. Each leaf speaks to the choice of having a place with a class of information confirming all tests way from the root to the leaf.

F. We utilize another order after choice tree which is guileless Bayes characterization the presentation of the classifiers is evaluated by contrasting it and other multi name. In the grouping calculation is applied by System to plan indicator that help acknowledgment of understudy's issues.

G. These results are given by stage 7 assistance teachers to recognize at issues understudies are confronting and settle on choices on appropriate obstruction to safeguard them and give better instruction framework.

IV. RESULTS:

Exactness, Precision and review are technique utilized for assessing the exhibition of supposition mining. Here exactness is the general precision of certain notion models. Review (Pos) and Precision (Pos) are the proportion and exactness proportion for genuine positive surveys. Review (Neg) and Precision (Neg) are the proportion and accuracy proportion for genuine negative surveys. In a perfect situation, all the exploratory outcomes are estimated by the Table 1.and exactness, Precision and review as clarified. below [9].

$$Recall := \frac{\#no.of\ correct\ outputs\ returned\ by\ the\ system}{\#no.of\ Total\ files\ Tested}$$

$$Precision := \frac{\#no.of\ Correct\ outputs\ returned\ by\ system}{\#no.Actual(true)predictions}$$

$$F1 - Measure := 2 * \frac{(R * p)}{R + P}$$

Evaluation of the system has been done using standard metrics Recall Precision and F1-Measure.

Table I: Accuracy comparison on Test Datasets.

Number of Experiments	Number of Reviews	Accuracy			
		Naïve Bayes (Movie Review)	KNN (Movie Review)	Naïve Bayes (Hotel Review)	KNN (Hotel Review)
1	100	57.18	48.13	42.12	45.78
2	200	63.23	55.78	41.78	41.00
3	500	71.16	59.12	43.89	41.34
4	1000	74.41	60.90	44.56	42.89
5	1500	77.53	64.87	48.45	45.65
6	2000	79.74	66.78	51.16	47.23
7	2500	80.12	67.67	52.00	48.33
8	3000	81.71	69.16	52.67	48.55
9	4000	82.54	69.34	53.97	49.67
10	4500	82.89	69.89	55.34	51.78

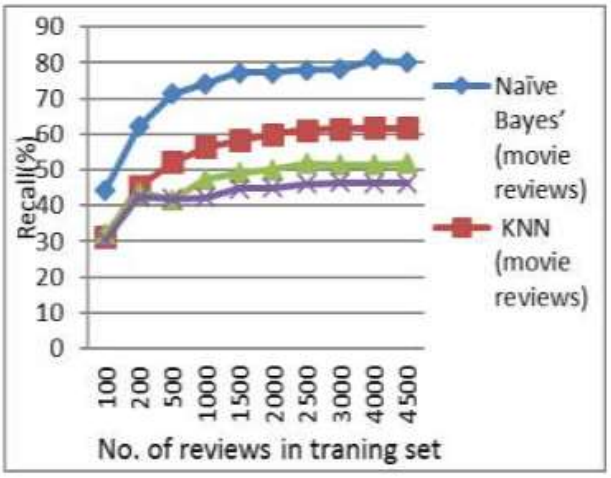


Figure 6: Recall comparison

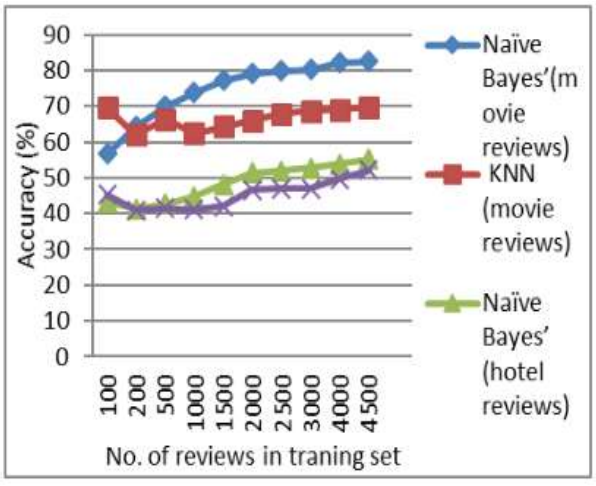


Figure 7: Accuracy comparison

Table II: Recall Comparison of datasets

Number of Experiments	Number of Reviews	Recall values			
		Naïve Bayes (Movie Review)	KNN (Movie Review)	Naïve Bayes (Hotel Review)	KNN (Hotel Review)
1	100	43.44	32.12	31.78	30.23
2	200	61.98	46.34	42.67	41.32
3	500	71.56	51.43	40.56	39.89
4	1000	73.98	55.97	48.01	41.45
5	1500	77.17	57.34	49.12	42.56
6	2000	77.56	60.98	49.89	45.67
7	2500	77.98	61.45	50.45	46.01
8	3000	78.60	61.67	51.22	46.32
9	4000	79.34	61.65	51.56	46.78
10	4500	80.14	61.89	51.98	46.88

V. CONCLUSION

The investigation covers how it identified the spam and non-spam messages utilizing Naïve Bayes calculation. This calculation permits taking care of huge number of highlights I. e. words. Subsequently, Naïve Bayes support for taking care of huge number words without any problem. Further, the informational index followed set of procedures of AI so as to construct the model. The informational collection procured from the Kaggle site and performed pre-handling utilizing numerous techniques including pack of-words. At that point, split the information into train and test model. Besides, it has assembled the model and assessed. The exactness of the model is tried utilizing exactness score, accuracy score, review score and F1 score.

REFERENCES:

[1] Belinkov, Y. and Glass, J. (2019). Analysis Methods in Neural Language Processing: A Survey.Transactions of the Association for Computational Linguistics, 7, pp.49-72.

[2] Cormack, G. (2008). Email Spam Filtering: A Systematic Review.Foundations and Trends® in Information Retrieval, 1(4), pp.335-455, DOI 10.1561/1500000006.

[3] Dada, E., Bassi, J., Chiroma, H., Abdulhamid, S., Adetunmbi, A. and Ajibuwa, O. (2019). Machine learning for email spam filtering: review, approaches and open research problems.Heliyon, 5(6), p. e01802.

[4] Dy, J.G and Broadley, C.E. (2004), Feature Selection for Unsupervised Learning, Journal of Machine Learning Research, 845–889. Available: http://www.jmlr.org/papers/volume5/dy04a/dy04a.pdf

[5] Edgar, T.W.,Manz, D. O. (2017),Using Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection. (2017).International Journal of Recent Trends in Engineering and Research, 3(4), pp.109-111.

[6] F.Y, O., J.E.T, A., O, A., J. O, H., O, O. and J, A. (2017). Supervised Machine Learning Algorithms: Classification and Comparison.International Journal of Computer Trends and Technology, 48(3), pp.128-138.

[7] Ghorbani, A., Steinhilber, G., Markus, D. and Maas, U. (2015). A PDF projection method: A pressure algorithm for stand-alone transported PDFs.Combustion Theory and Modelling, 19(2), pp.188-222.

[8] Khanum, M., Mahboob, T., Imtiaz, W., Abdul Ghafoor, H. and Sehar, R. (2015). A Survey on Unsupervised Machine Learning Algorithms for Automation, Classification and Maintenance.International Journal of Computer Applications, 119(13), pp.34-39.

[9] Leimbach, M. (1994). Expert system model coupling within the framework of an ecological advisory system.Ecological Modelling, 75-76, pp.589-600.

[10] Mao, H., Alizadeh, M., Menache, I. and Kandula, S. (2016), Resource Management with Deep Reinforcement Learning. In ACM Workshop onHot Topics in Networks.

[11] Prakash, V. and Nithya, D. (2014). A Survey On Semi-Supervised Learning Techniques.International Journal of Computer Trends and Technology, 8(1), pp.29.

[12] Omar, S., Ngadi, A. and H. Jebur, H. (2013). Machine Learning Techniques for Anomaly Detection: An Overview.International Journal of Computer Applications, 79(2), pp.33-41.

[13] Rennie, J. D., Shih, L., Teevan, J., & Karger, D. R. (2003). Tackling the poor assumptions of naive Bayestext classifiers.In ICML (Vol. 3, pp. 616-623).[14] Wang, H. (2013). Quality Measurements for Association Rules Hiding.AASRI Procedia, 5, pp.228-234.