Sentiment Analysis based Method for Amazon Product Reviews

Ms. Jyoti Budhwar
Department of Computer Science & Engineering
Deenbandhu Chottu Ram University of Science & Technology, Murthal

Prof. Sukhdip Singh
Department of Computer Science & Engineering
Deenbandhu Chottu Ram University of Science & Technology, Murthal

Abstract—**Research is focusing to apply sentiment analysis to review the product of Amazon. Research is using hybrid approach that is making use of Naïve bayes approach, KNN, and LSTM mechanism. Naïve bayes provided solution for classification. And KNN helps in grouping. The data set would be trained using LSTM based model to provide more accuracy in solution. Data set of review of customer has been considered in order to perform sentiment analysis. The proposed research is supposed to resolve the issues of previous research that were faced during sentiment analysis.**

Keywords— Sentiment analysis, KNN, LSTM, Naïve bayes

I. INTRODUCTION

Reviewing product using sentiment analysis is becoming popular for text mining. Research is also considering research in area of computational linguistics. Research work is focusing on correlation among Amazon product reviews. Research is also considering rating of products provided by customers. Research has considered traditional machine learning algorithms along with Naive Bayes analysis, SVM, Knearest neighbor mechanism. Research has also considered deep neural networks along with Recurrent Neural Network (RNN). Research is supposed to provide better solution for sentiment analysis.

Fig 1 Customer behavior during Sentiment Analysis [9]

There is increment in amount of research efforts related to sentiment in textual resources in recent years. The researches published on the sentiment analysis are increasing in last years. Those research works have been considered in present research paper. Research work is also dealing with sub topic that has been known as sentiment analysis. This is also known as opinion mining. This has been presented as group of textual content. Such researches consider the opinion of people, appraisals, attitudes, and emotions for entities. Moreover the issues, individuals, events, concept with their features are considered. The application build using such type of concept has diverse nature.

In today's world, any company must take customer feedback into account. Customers' feelings are taken into account when designing goods and services. Before using a programme or purchasing a product, potential consumers consider the thoughts and feelings of current users. Furthermore, researcher [2] uses this data to do an in-depth study of industry dynamics and customer preferences. These type of opinions could lead to good forecasting in stock market.

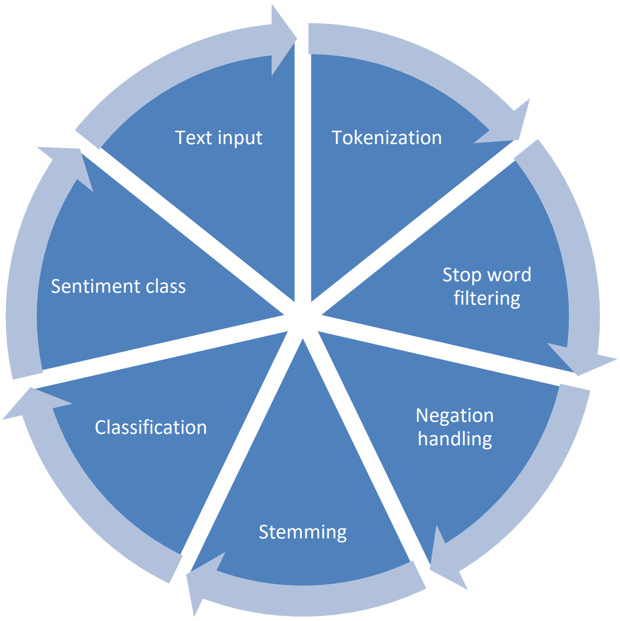

Fig 2 Sentiment analysis approach for prediction [10]

Despite this, finding and monitoring online opinion pages, as well as distilling the facts found in them, remains a difficult challenge due to the abundance of different sites. In long forum postings and blogs, each site usually includes a large amount of opinionated text that is not always easy to decipher.

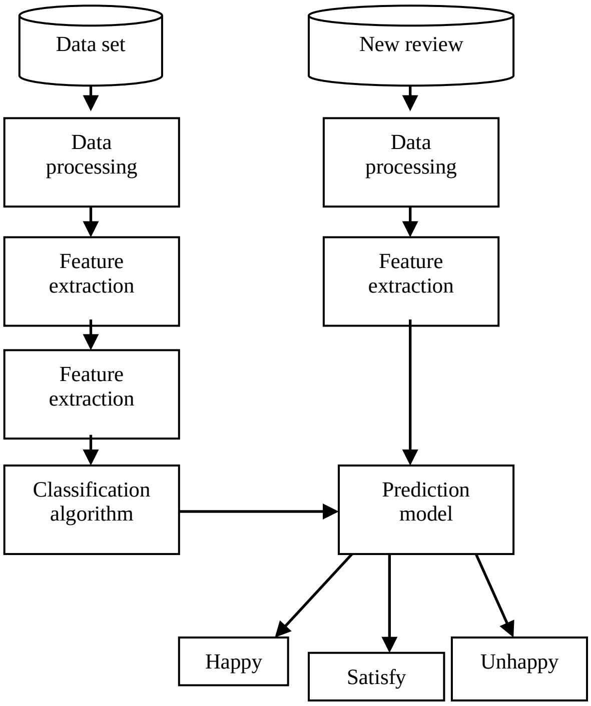

Fig 3 Prediction model for sentiment analysis [11]

II. LITERATURE REVIEW

There have been several academic papers published so far on product ratings, sentiment analysis, and opinion mining. On Yelp's ranking dataset, for example, Xu Yun[8] et al from Stanford University used existing supervised learning algorithms like the perceptron algorithm, naive bayes, and supporting vector machine to predict a review's rank. They carried out cross validation with 70% of the data.

Maria Soledad Elli [3] did sentiment from considering reviews of customer. They have analyzed result to develop model for business. Author presented that tool is providing better accuracy. Research have make use of Multinomial Naive Bayesian. It is acting as classifiers. Mechanism is also supporting vector machine.

Callen Rain [6] has extended research work in area of processing of natural language. Research is making use of Naive Bayesian as well as decision list classifiers. These mechanism have been utilized to categorize a provided review. These review could be positive or negative. Research is making use of Deep-learning neural networks. Neural network has been found famous in field of sentiment analysis.

Ronan Collobert [1] et al has made use of convolutional network. Semantic role labeling task has been performed with the objective of avoiding too much operation oriented engineering of characteristics.

On the other hand, in paper [7], the authors proposed using recursive neural networks to achieve a better understanding compositionality in tasks such as sentiment detection. In this paper, we want to apply both traditional algorithms including Naive Bayesian, K-nearest neighbor, Supporting Vector Machine and deep-learning tricks.

Table 1 Existing Researches

Sno.	Author / Year	Objective of research	Methodology	Limitation
1.	R. Collobert / 2011	To implement Natural language processing from scratch.	Machine learning	Research failed to provide solution for semantic review
2.	K. Dave/ 2003	Performing Opinion extraction and semantic classification of product reviews.	Semantic classification	Research has not considered optimized solution.
3.	M. S. Elli	To propose Amazon reviews, business analytics with sentiment analysis.	Analyzing the Sentiment	There is lack of accuracy is prediction
4.	S. Hota / 2018	Proposing Knn classifier based approach for multi-class sentiment analysis of twitter data.	Knn classifier	This work is suffering from performance issues.
5.	B. Liu / 2012	To perform Opinion Mining and Sentiment Analysis	Sentiment analysis	There is lack of accuracy and flexibility.
6.	C. Rain. / 2013	Implementing Sentiment analysis in amazon reviews using probabilistic machine learning.	Machine learning	Research is providing solution on the basis of probability that leads to degradation in accuracy.
7.	R. Socher/ 2013	Presenting Recursive deep models for semantic compositionality over a sentiment treebank.	Recursive deep model	Recursive deep model wastes lot of time during training.
8.	Y. Xu / 2015	Implementing Sentiment analysis of yelps ratings based on text reviews	Sentiment analysis	Research is not provided wide scope.

III. PROBLEM STATEMENT

However there have been several researches in field of sentiment analysis. But existing research have provided limited scope. Moreover the performance factor during sentiment analysis is ignored. The neural network model used in previous research is taking lot of time during processing. Several researches are providing solution on the basis of probability that leads to degradation in accuracy. There is need to provide solution that consider traditional machine learning algorithms along with Naive Bayes analysis, SVM, Knearest neighbor mechanism. Research is supposed to consider deep neural networks along with Recurrent Neural Network (RNN). Research is supposed to provide better solution for sentiment analysis.

IV. DATASET AND FEATURES

A. Data Preprocessing

Data is already collected out of assessment records which has been done by customer. This assessment is done on the products which are provided by Amazon . Near about thirty five thousand informational points are enclosed by such type of data . Each example includes the type, name of the product as well as the text review and the rating of the product. For the utilization of data in a refine way, two, most important column related to this project are extracted by us initially. These two columns are rating and assessment. After that, when the data is checked by us, we noticed those information points which remain unrated. When we eliminate such sample, we generally left with thirty four thousand and six hundred twenty seven information points. On the other hand , for outlining the data, allocation of ratings has been plotted by us. There have been 5 classes. Rating from 1 to 5 has been distributed in the middle of these classes. In reality, such type of classes are uneven. The basic reason behind this unevenness is the availability of less number of information in class one as well as in two. At the same time, twenty thousand reviews is possessed by class five . During research review text has been converted into an input vector.

B. Data Resampling

Further sampling of data is needed by us in support of our samples because of unevenness of data. Additional sampling of data becomes the most famous method for dealing in the company of uneven data. Here, the information related to class one, two and three is over scanned by us. The basic reason behind the over scanning of these classes is the availability of less samples in comparison to remaining classes. This is the reason, due to which, initial analysis related to label one , two and three came fifteen times in those studies which are established by us. On the other hand, availability of repeated samples make the design over fit .
Here, in this academic study usual methods are used by the scholars. Generally, a thesaurus is established by us on the basis of usual term and arrange usual term.
Boundaries in support of term thesaurus has been come in six occurrence. It eventually collected four thousand two hundred and twenty three terms out of ovelall dataset. After that, all the reviews are converted into a table. Here, appearance of each term is represented by each value. Modification ion the threshold and the length of the dictionary is opted. It becomes important to note that expansion of dictionary's extent fails to put considerable effect on precesion.

V. TOOLS AND TECHNIQUES

Research has used hybrid approach which is integration of Naïve bayes approach, KNN, and LSTM. Naïve bayes is classifying dataset whereas KNN helps in grouping. The data set has been trained with the support of LSTM based model in order to increase accuracy. Data set of review of customer would be considered to perform sentiment analysis.

A. 5.1. Naive Bayes

It becomes the most famous and productive training method at the time of rating problems. It is assumed by this method that x 0 become independent in certain condition. It become famous in the form of Naive Bayes assumption.

$$p(x_1, ..., x_k|y) = \prod_{i=1}^{k} p(x_i|y)$$

For improving the working of our design Laplace Smoothing is also integrated by us. On the basis of formula given below a sample is forecasted:

$$\hat{y}^{(i)} = \arg\max_j \prod_{i=1}^{k} p(x_i|y = j)\phi(j)$$

Intianlly, for representing the review of text material, it need an arrangement of favourable integers, and designs p(xi |y) in the company of general allocation. With the second way of representing review texts using glove dictionary, the inputs fails to remain favourable integers, so we chose to model p(xi |y) in the company of Gaussian allocation.

B. K-nearest Neighbor

It becomes famous in the form of statistic rating method. In recent years, it has been used extensively . At the time of prediction, it search out for K is equal to n. After that, the major part of such neighbours' is assigned by it. The distance in the middle of adjacent neighbour becomes famous in the form of euclidean distance. It can easily determine the similar point of each dataset.[4]

$$\hat{f}(x) = \frac{1}{K} \sum_{x \in N_K(x)} y_i$$

The arithmetic form of this method is shown by the above equation . The usual concept of this methods is that when inputs are identically interconnected , then the output are also identical. Here, amount of K is tuned by us in the middle of four five and six .

C. Linear Support Vector Machine

It becomes famous in the form of method which builds an organizer which isolate the marked data. Geometrically given two types of points, circles and x's, in a space, it tries to maximize the minimum distance from one of the points to the other. This means that, margin is maximized by it . This method attempted to figure out maximization problem which is given below:

$$\arg\max_{\gamma,w,b} \frac{1}{2}||w||^2$$
$$s.t. y^i(w^T x + b) \geq 1, i = 1, 2, ...m$$

For satisfying maximum margin problem and separability constraint, it has been determined w

D. Long Short Term Memory

It exists in the form of RNN section. An usual long short term memory section is made from unit, input , output gate and forget gate. Unit memorize values for unpredictable duration. and the three gates regulate the flow of information into and out of the cell. Its networks become appropriate for the classification, processing and making predictions based on time series data, since there can be lags of unknown time period in the middle of significant events in a time series. Structure is represented in the next figure.

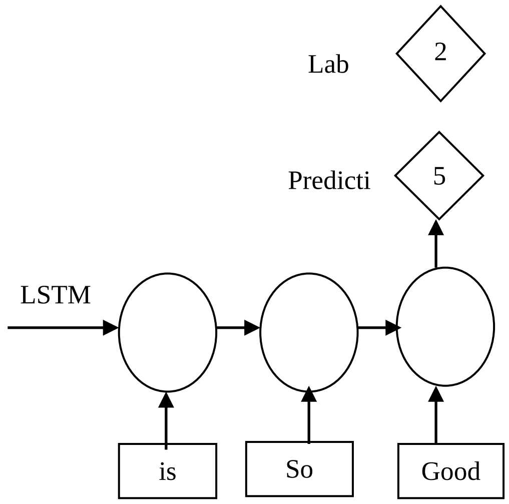

Fig 4 LSTM Configuration

LSTM is with support of special units in addition to standard units. LSTM units are consisting of a 'memory cell'. These memory cells are capable to maintain data in memory for large time. Users are moving from **RNN** to **LSTM** because it is introducing more controlling knobs. They are capable to manage flow and mixing of Inputs according to trained Weights. So it provides flexibility during management of outputs. Thus **LSTM** is providing ability to manage and good results.

VI. CONCLUSION AND FUTURE WORK

In summary, proposed work tried Naive Bayes, SVM, KNN, LSTM. Research is supposed to provide more flexible and accurate solution. Naïve bayes provided solution for classification. And KNN helps in grouping. The data set would be trained using LSTM based model to provide more accuracy in solution. The proposed research is supposed to resolve the issue of previous research that was faced during sentiment analysis.

ACKNOWLEDGMENT

I am thankful to my guide who has assisted be in preparing research paper. I hope that this paper would also assist other researchers to get better solution in field of sentiment analysis.

REFERENCES

[1] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. Natural language processing (almost) from scratch. Journal of Machine Learning Research, 12(Aug):2493–2537, 2011.

[2] K. Dave, S. Lawrence, and D. M. Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In Proceedings of the 12th international conference on World Wide Web, pages 519–528. ACM, 2003.

[3] M. S. Elli and Y.-F. Wang. Amazon reviews, business analytics with sentiment analysis.

[4] S. Hota and S. Pathak. Knn classifier based approach for multi-class sentiment analysis of twitter data. In International Journal of Engineering Technology, pages 1372–1375. SPC, 2018.

[5] B. Liu and L. Zhang. A Survey of Opinion Mining and Sentiment Analysis, pages 415–463. Springer US, Boston, MA, 2012.

[6] C. Rain. Sentiment analysis in amazon reviews using probabilistic machine learning. Swarthmore College, 2013.

[7] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the 2013 conference on empirical methods in natural language processing, pages 1631–1642, 2013.

[8] Y. Xu, X. Wu, and Q. Wang. Sentiment analysis of yelps ratings based on text reviews, 2015.

[9] https://mk0ecommercefas531pc.kinstacdn.com/wp-content/uploads/2019/12/sentiment-analysis.png

[10] https://miro.medium.com/max/361/0*ga5rNPmVYBsCm-lz.

[11] https://i.ytimg.com/vi/VXt9SQx5eM0/maxresdefault.jpg