2021 Fourth International Conference on Computational Intelligence and Communication Technologies (CCICT) | 978-1-6654-2392-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/CCICT53244.2021.00033

Sentiment Analysis of Amazon Food Review Data

Sneha Choudhary
Assistant Professor, JIMS
Snehachoudhary3012@gmail.com

Charu Chhabra
Assistant Professor, RDIAS, GGSIPU
Corporate.charu@gmail.com

Abstract

Sentiment classification is a functioning examination area in information mining and information disclosure with different application spaces. Accomplishment of item advancement sites, for example, Amazon, eBay and so forth gets influenced by the nature of the surveys they have for their yields. Every one of these destinations gives a path to the analyst to form the remarks on the basis of the items and assign a remark to it. Considering these remarks, the analysis will be classified as best or worst. By this, a structure can be edified that can identify the sentiment masked in a review, performing sentiment categorization on a gigantic dataset. All particulars can be grouped into primarily two classes, facts and opinions. Facts are assertions about matter and worldly occurrences. And opinions are individual statements that mirror individuals' assumptions or bits of knowledge about the entities and events. This paper shows the performance of classification algorithms such as Decision Tree, Bernoulli NB, Logistic, and Perceptron using Principal Component Analysis (PCA), applying n-gram (unigram, bigram) on the entire feature set and computing confusion matrix for the dataset.

Keywords

Sentiment Classification, Decision Tree, Principal Component Analysis, Logistic Based Classification, Perceptron classification, N-gram Confusion Matrix

I. INTRODUCTION

The Internet is known as the most popular correspondence stage for open surveys, feelings, remarks and conclusions. The number of dynamic clients and the size of their audits made day by day on online sites are colossal. There are 2.4 billion dynamic online clients, who compose and read online around the globe. According to a 2013 Study, 79% of customer's self-assurance is based on online personal reference reviews. Amazon is one of the biggest online shippers on the planet. Individuals frequently look over the items before buying the item on the website. But the reviews on amazon are not essentially products but a combination of product review and service remarks. In this paper, we create a new proposed method for evaluating methodical papers based on sentiment analysis and domain factors. Sentiment analysis targets characterizing the mentality of a client as for certain subjects or the total feeling of a text, for example, positive or negative. It relies upon two issues, opinion extremity and estimation score and is a parallel worth whichever positive or negative. The purchaser is beguiled as the general sentiment (rating classification) that amazon gives is an assembled one and there is no distinction between the reviews. The anticipated model appropriately detaches service and product review, notwithstanding this. It likewise orders the review as Feature review if the client discusses some particular item highlight. A highlighted review is only an item survey; our model likewise gives estimation of the original copy about the item included. We expect to manufacture a framework that photos the reviewer's opinion as diagrams.

II. METHODOLOGY

A. Classification Algorithm

- **Assignment:** Regulate which of a static set of classes an instance belongs to
- **Contribution:** Training set of examples annotated with class standards.
- **Result** Induced guess (model/concept description/classifiers)
- **Learning** Persuade classifiers from training data
- **Prediction** Using Theory for Prediction

B. Logistic Regression

Logistic regression is a mechanism for modelling a binary dependent element, which has values 1 and 0. The logit function is used to change a curve into a straight line and the range of the proportion changes from 0–1 to $-\infty$ to $+\infty$.

C. Decision Tree

Decision Tree is the easiest presentation for classifying patterns. It is a Supervised Machine Learning where the data is incessantly split on the basis of certain criteria. Decision Tree consists of the following components namely Nodes, Branches and Leaf Nodes.

D. Bernoulli Naive Bayes

A Naive Bayes classifier is a likelihood machine learning model that is used for categorization activities. The kernel of the categorizer is based on the Bayes theorem.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

978-1-6654-2392-2/21/$31.00 ©2021 IEEE
DOI 10.1109/CCICT53244.2021.00033

Using Bayes theorem, we can find the probability of A given that B has transpired. Here, B is the evidence and A is the hypothesis. The supposition is that the features are independent. That is why it is known as naive. The predictors are Boolean variables and the criterion that we use to hypothesize the class variable takes only yes or no.

E. Perceptron

Perceptron is a function for binary criterion that uses a linear prediction function:

f(x) = 1, wTx+ b ≥ 0-1, wTx+ b < 0

This is called a step function, which interprets the result as 1 if "**wTx+ b ≥0**" is true, and the resultant is -1 if rather "wTx+ b < 0" is true.

III. Feature Reduction using Principal Component Analysis (PCA)

Dimensionality reduction is a way to reduce the entanglement of a model and avoid overfitting. There are two main types of dimensionality reduction: feature selection and feature extraction. In feature selection, we choose a subset of the original aspect, whereas in feature extraction, we acquire information from the feature set to design a new feature subspace. Principal Component Analysis (PCA) algorithm is used to constrict a dataset into a lower-magnitude feature subspace with most of the pertinent knowledge. This paper aims to make an intensive study of the effectiveness of reduced features using Principal Component Analysis (PCA) for sentiment classification of tasks of online product reviews. Also, we are only finding remarks for food evaluation. The effectiveness of the features thus selected is evaluated using Decision tree, Perceptron, Logistic and Bernoulli Naive Bayes classifier.

IV. Implementation and Working Model

1. Reading the data from the SQLite file.
2. Encoding score to Positive or negative based on value of each sample.
3. Distribution of labels in the dataset.
4. Splitting the dataset based on labels.
5. Preprocessing.
6. Splitting the data into train and test.
7. Feature Reduction & Selection
8. Trying n-gram (unigram, bigram) on the entire feature set.
9. Compute confusion matrix.

V. Data collection and Analysis

The Amazon Fine Food Reviews dataset is a 300 megabytes huge dataset which comprises approximately 568k mentions of amazon food items between the years 1999 and 2012. In Fig 1 the flowchart of sentiment analysis is depicted.

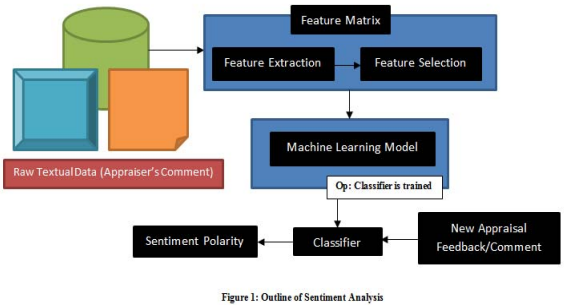

Figure 1: Outline of Sentiment Analysis

Fig 1: Outline of Sentiment Analysis

VI. Result and Discussions

A. Distribution of labels in dataset

The figure 2 depicts the positive negative score

Score:
Negative 124677
Positive 443777

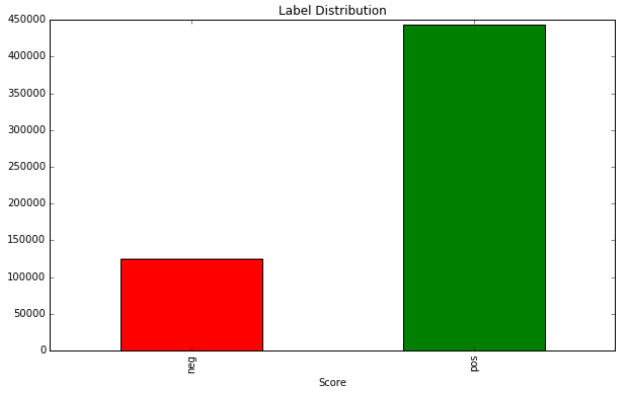

Figure 2: Positive Negative score

Percentage of negative reviews 21.93 %
Percentage of positive reviews 78.07 %
The mean of scores is 4.1.

B. Using PCA:

The feature set is minimized to 200 workings using abbreviated Singular Value Decomposition and works on scarce matrices. Figure 3 depicts the accuracy of classifiers. Figure 4 depicts the bar graph accuracy of classifiers , feature selection: (On frequent words). Fig 5 represents accuracy of classifiers on frequent words and Fig 6 depicts bar graph accuracy of classifiers on frequent words

	Model	TF-IDF Accuracy
0	Decision Tree	0.858473
1	BernoulliNB	0.793455
2	Logistic	0.844667
3	Perceptron	0.814593

Figure 3: Accuracy of classifiers

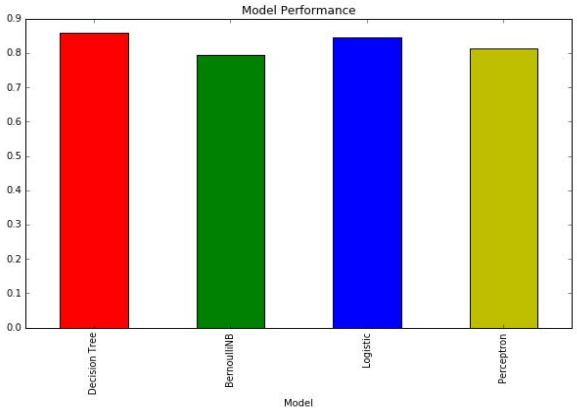

Fig 4: Bar graph accuracy of classifiers

Feature selection:
(On frequent words):

	TF-IDF Accuracy	Model
0	0.915047	Decision Tree
1	0.869049	BernoulliNB
2	0.888955	Logistic
3	0.851753	Perceptron

Fig 5: Accuracy of classifiers on frequent words

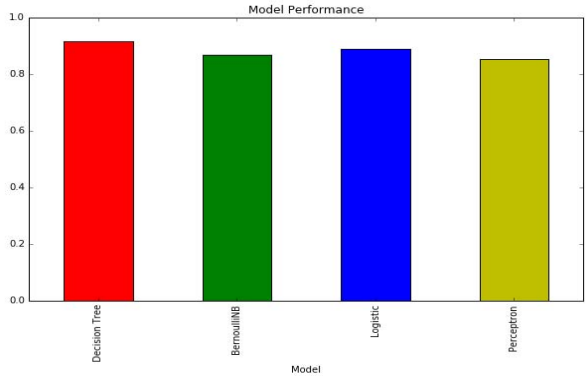

Fig 6: Bar graph accuracy of classifiers on frequent words

C. Term Frequency- Inverse Document Frequency

TF-IDF is a set of rules that counts the word weight by keeping in mind the frequency of the word (TF) and in how many documents the word can be found (IDF). Fig 7 depicts a histogram of TF-IDF on frequent words.

TF(t) = (Number of times term t appears in a document) / (Total number of terms in the document)
IDF(t) = loge (Total number of documents / Number of documents with term t in it)

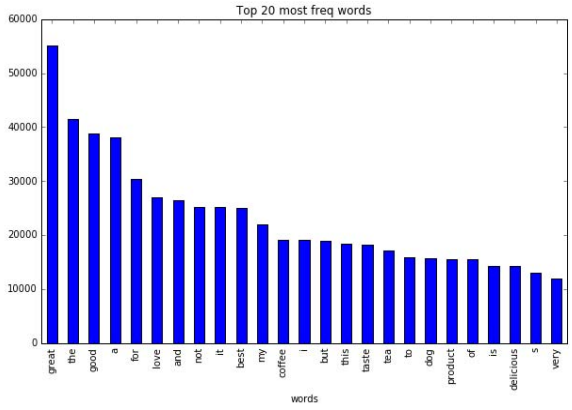

Fig 7: Histogram of TF-IDF on frequent words

D. N-Gram Models on entire feature set

We can make use of a cluster of words to yield better results. Sometimes the order of words may have an unusual effect on the prediction. Sequences like "not good" or "not bad" affect the likelihood in ways dissimilar than when used individually.Fig 8(A) represents Accuracy of Unigram model and Fig 8(B) represents Accuracy of Bigram model

A. Unigram: **B. Bigram:**

	TF-IDF Accuracy	Model
0	0.887231	BernoulliNB
1	0.928550	Logistic
2	0.903760	Perceptron

Fig 8(A): Accuracy of Unigram model

	TF-IDF Accuracy	Model
0	0.873644	BernoulliNB
1	0.893339	Logistic
2	0.861090	Perceptron

Fig 8(B): Accuracy of Bigram model

E. Compute Confusion Matrix

From the matrix it is clear that many samples were anticipated to be positive and their actual label was also positive. Also, few negative samples which were predicted negative also came out to be truly negative. But this matrix is not suggestive of the performance because in testing data the negative samples were very less, so it is likely to see the predicted label vs true label part of the matrix for negative labels as lightly shaded. Fig 10 depicts Confusion Matrix

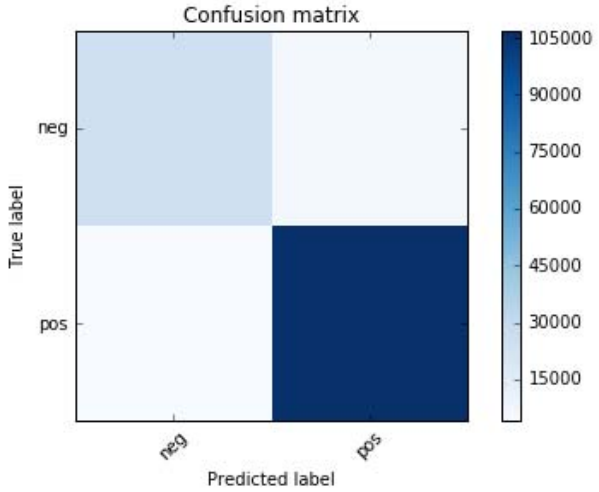

Fig 9: Confusion Matrix

VII. CONCLUSION

The main focus is to confirm unbiased results of sentiments, in order to reduce the time complexity, summarization of the results in the form of charts is done (Statistical Graphs). Information representation is a significant innovation in the coming future, as information is expanding in size and multifaceted nature. Consequently, our framework sums up the outcomes as bar outlines and pie graphs that help clients to see and straightforwardly comprehend the notion separated. Our model is grouping the audits and doing a slant examination on it. Achievement of thing-based merchandise/items selling sites, for example, Amazon, eBay and so on gets hampered by the nature of the surveys they have for their items. Every one of these destinations gives a path to the commentator to compose his/her remarks about the administration or item and give a rating for it. In light of these remarks one can arrange each audit as best or worst. From this information, a model can be capable of perceiving the conclusion concealed in a survey, doing feeling characterization on an immense dataset.

VIII. REFERENCES

[1] "Statistics review 14: Logistic regression | SpringerLink." 13 Jan. 2005, https://link.springer.com/article/10.1186/cc3045. Accessed 19 Aug. 2020.

[2] "Decision Tree Classification. A Decision Tree is a simple ...." 5 Jul. 2019, https://towardsdatascience.com/decision-tree-classification-de64fc4d5aac. Accessed 19 Aug. 2020.

[3] "Naive Bayes Classifier. What is a classifier? | by Rohith ...." 5 May. 2018, https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c. Accessed 19 Aug. 2020.

[4] "Principal Component Analysis for Dimensionality Reduction ...." 24 May. 2019, https://towardsdatascience.com/principal-component-analysis-for-dimensionality-reduction-115a3d157bad. Accessed 19 Aug. 2020.

[5] "Sentiment Analysis for Appraisal Data |." 23 Apr. 2014, https://softwareunwound.wordpress.com/2014/04/23/sentiment-analysis-for-appraisal-data/. Accessed 19 Aug. 2020.

[6] "Automated document classification for news articles in Bahasa ...." https://ieeexplore.ieee.org/iel7/6996670/7006983/07007894.pdf. Accessed 19 Aug. 2020.

[7] Zhang, X., & Zheng, X. (2016). Comparison of Text Sentiment Analysis Based on Machine Learning. 2016 15th International Symposium on Parallel and Distributed Computing (ISPDC). https://doi.org/10.1109/ispdc.2016.39

[8] Haberzettl, M., & Markscheffel, B. (2018). A Literature Analysis for the Identification of Machine Learning and Feature Extraction Methods for Sentiment Analysis. *2018 Thirteenth International Conference on Digital Information Management (ICDIM)*. https://doi.org/10.1109/icdim.2018.8846980

[9] S., H., & Ramathmika, R. (2019). Sentiment Analysis of Yelp Reviews by Machine Learning. 2019 International Conference on Intelligent Computing and Control Systems (ICCS). https://doi.org/10.1109/iccs45141.2019.9065812.

[10] Niu, W., & Wu, L. (2019). Sentiment Analysis and Contrastive Experiments of Long News Texts. 2019 IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC). https://doi.org/10.1109/iaeac47372.2019.8997550

[11] Woldemariam, Y. (2016). Sentiment analysis in a cross-media analysis framework. 2016 IEEE International Conference on Big Data Analysis (ICBDA). https://doi.org/10.1109/icbda.2016.7509790

[12] Chaturvedi, S., Mishra, V., & Mishra, N. (2017). Sentiment analysis using machine learning for business intelligence. 2017 IEEE

International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI). https://doi.org/10.1109/icpcsi.2017.8392100
[13] Rathi, M., Malik, A., Varshney, D., Sharma, R., & Mendiratta, S. (2018). Sentiment Analysis of Tweets Using Machine Learning Approach. 2018 Eleventh International Conference on Contemporary Computing (IC3). https://doi.org/10.1109/ic3.2018.8530517
[14] Zharmagambetov, A. S., & Pak, A. A. (2015). Sentiment analysis of a document using deep learning approach and decision trees. 2015 Twelve International Conference on Electronics Computer and Computation (ICECCO). https://doi.org/10.1109/icecco.2015.7416902
[15] Daniati, E., & Utama, H. (2020). Decision Making Framework Based On Sentiment Analysis in Twitter Using SAW and Machine Learning Approach. 2020 3rd International Conference on Information and Communications Technology (ICOIACT). https://doi.org/10.1109/icoiact50329.2020.9331998
[16] Yadav, Y., Kumar, V., Ranga, V., & Rawat, R. M. (2020). Analysis of Facial Sentiments: A deep-learning Way. 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC). https://doi.org/10.1109/icesc48915.2020.9155622
[17] Hegde, R., & Seema S. (2017). Aspect based feature extraction and sentiment classification of review data sets using Incremental machine learning algorithm. 2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB). https://doi.org/10.1109/aeeicb.2017.7972395
[18] Seals, E., & Price, S. R. (2020). Preliminary Investigation in the use of Sentiment Analysis in Prediction of Stock Forecasting using Machine Learning. 2020 SoutheastCon. https://doi.org/10.1109/southeastcon44009.2020.9368258
[19] Sajib, M. I., Mahmud Shargo, S., & Hossain, M. A. (2019). Comparison of the efficiency of Machine Learning algorithms on Twitter Sentiment Analysis of Pathao. 2019 22nd International Conference on Computer and Information Technology (ICCIT). https://doi.org/10.1109/iccit48885.2019.9038208
[20] Juneja, P., & Ojha, U. (2017). Casting online votes: To predict offline results using sentiment analysis by machine learning classifiers. 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT). https://doi.org/10.1109/icccnt.2017.8203996