IDENTIFY THE REVIEW IS FAKE OR ORIGINAL THROUGH SENTIMENT ANALYSIS USING ML

[1]D. SRIKANTH, DR. B. SAI JYOTHI[2]

[1]MCA STUDENT, DEPARTMENT OF MASTER OF COMPUTER APPLICATIONS VASIREDDY VENKATADRI INSTITUTE OF TECHNOLOGY(VVIT), NAMBURU, GUNTUR, AP.

[2]PROFESSOR, DEPARTMENT OF INFORMATION TECHNOLOGY VASIREDDY VENKATADRI INSTITUTE OF TECHNOLOGY(VVIT),

Abstract**: - Sentiment analysis is also called as opinion mining. Sentiment analysis is widely applied to voice of the customer materials such as reviews and survey responses. This project aims to classify a particular item reviews into bunch of positive or negative polarity by utilizing machine learning algorithms. [1] Sentiment Analysis and text classification methods are applied to a dataset of a particular product reviews. Additionally, system compares two supervised machine learning algorithms are Support Vector Machine(SVM)[3], Decision Tree(DT-J48) for sentiment classification of reviews utilizing these two different datasets[5]. Including item reviews on dataset amazon.**

Existing system uses sentiment analysis method in order to detect the fake reviews. Proposed system detects the fake reviews, gender specific fake reviews and [4] accuracy of algorithms. Final result of the project shows Support Vector Machine algorithm is better than using of Decision Tree(DT-j48) algorithm.

Keywords: Sentiment Analysis, Fake reviews, Support Vector Machine, Decision Tree-J48.

I. INTRODUCTION

The purpose of the project is to detect fake reviews and find out which algorithm is the best with the highest precision. With the help of reviews, the customers buy the best online products. After using the product, the customer writes the review then it helps to make the right decision. In addition, display the fake or original review and the fake review percentage and fake percentage analysis and fake [13] gender-specific analysis on the output panel. People who are buying online products using this project that type of customers. They refer to the reviews prior to purchasing the product. This project is of use to buyers online.

[7] Text processing is the automated method of unstructured text comprehension and sorting, making it easier to handle. For example, Word cloud tools are used to perform very basic text analysis techniques, such as detecting keywords and phrases that most often appear in your data. Most people need a product overview before their money is spent on the product. So people come across different reviews in the website but the user does not identify these reviews as original or fake. Some good reviews are added by the product company people themselves in some review websites

in order to make the product famous these people belong to the Social Media Optimization team. They give good reviews for many different products their own company manufactures. Users will not be able to determine if the review is original or fake. The goal to identify fake reviews by using the method of machine learning to identify the original reviews by sentiment analysis. Customers will be encouraged to buy more online items by removing false reviews and preserving original reviews. In this project, the SVM and DT-J48 supervised machine learning algorithms are used. Using the Verified Purchase column of a data set, the reviews are categorized into false and original and the revision show is an original or fake application of ML techniques and sentiment classification algorithms to break the dataset into a qualified and test dataset.

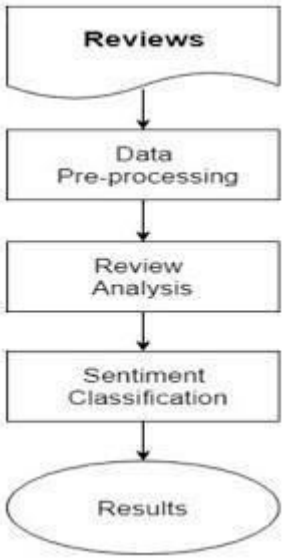

Fig 1. *Sentiment Analysis*

Users may provide the input analysis and the output is original or false. Supporting Vector Machine(SVM) and Decisions Tree (DT-j48) are the two supervised machine learning algorithms. Show the algorithms accuracy and potential gender specification. Find out, in addition to the text classification, which algorithm is the best in the highest accuracy. Customer feedback is a customer comment form on e-commerce and online purchase sites. Checks help customers purchase the best product on websites for e-commerce. The customer can sometimes make false reviews. The positive false reviews and the negative counterfeit reviews are classified.

The customer is compelled to take action in positive false reviews and the customer is persuaded no longer to take action at the product being reviewed in negative false reviews. It is written to damage the reputation of one company. Why the fake review is written as Emotions, Funding, Concourses, Established facility

II. RELATED WORKS

Our analysis uses statistical methods for evaluating the Detection mechanism performance for fake reviews, and evaluate that detection accuracy. Hence, the current program examination of the literature on the research using statistical approaches.

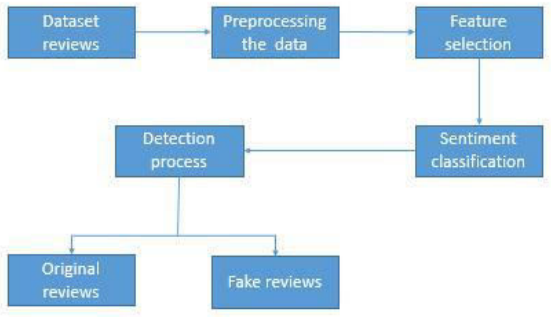

Fig 2. *Design of the system*

A. Sentiment analysis

When carrying out SA[6], several problems are to be considered. Two major issues are addressed in this section. Firstly, in another situation, the viewpoint (or opinion) observed as negative may be considered positive. Second, people don't always express similar opinions. But importantly

Popular text processing techniques employ the assumption that small modifications are unlikely to change the actual meaning between the two text fragments.

B. Detecting Fake reviews using Machine Learning:

The following stages are the machine learning process. First, there is a training dataset, that can be marked with sentiment labels or not. Second, each review is shown as a vector. Third, by analyzing relevant features, a classifier is trained to differentiate between sentiment labels. Finally, the trained classification is used to identify new reviews. The module LabelEncoder is used to preprocess the text and divide the dataset as trained and test. The text translates as 0 and 1. 0 is considered fake in the Verified Purchase column dataset, and 1 is considered as real.

C. Textual reviews:

Product reviews are based on an increase in sales or a growth in consumer purchases for an e-commerce store. In order to increase sales by informing consumers about how to make the decision to buy the product, an online review is very important for companies. The products recommended by other users are always more likely to be purchased by people.

D. Gender specification:

Review text is translated to binary code in the dataset. A binary code represents text, instructions to the computer processor, or any other data that uses a two-symbol system. The two-symbol system used on the binary number system is always "0" and "1." The binary code assigns every character, instruction, etc. Binary digit patterns also called bits. Through the dictionary concept, that binary code is compared to [14] gender and verified purchasing columns within the dataset.

E. Comparative study of different Classification algorithms

Analysis of review through the sentiment analysis using ML techniques. The two supervised machine learning algorithms are SVM and DT-j48 are applied to the dataset. Using the metrics module to calculate the accuracy score of the two algorithms. [4] Based on the accuracy score to decide the which is best algorithm as shown below the table.

TABLE 1. COMPARISON OF ACCURACY OF CLASSIFIERS

Classification algorithms	**Accuracy %**
Support Vector Machine	0.64
Decision Tree	54.6

III. METHODOLOGY

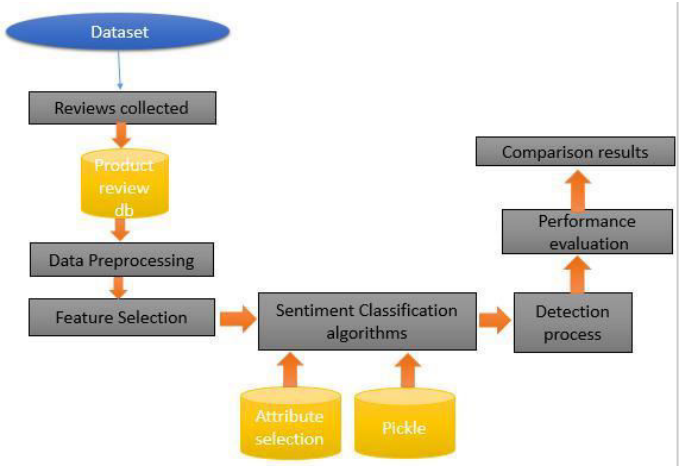

Fig 3. *System Architecture*

Step 1: Dataset

Sample dataset stores the following information.

- PRODUCT_TITLE: Number of products are available on websites like Amazon for e-commerce purposes. Selected objects are included in the sample data collection.
- VERIFIED PURCHASE: This dataset column is used to determine whether or not the customer actually buys the product. This is used to find those who write fake reviews.
- REVIEW TEXT: Consumers write product reviews containing product benefits and drawbacks as well as recommendations for potential applications.
- GENDER: Store consumer gender in this section. This column is useful for the user interface gender.

Step 2: Reviwes collection

The experiment is based on an analysis of the sentiment value of the standards dataset in order to provide an exhaustive study of machine learning algorithms. This system used the amazon reviews' original dataset for testing the classification of our review methods. The data set is now available and is used to store information that is often granted for researchers in the field of sentiment analysis as a standard gold dataset. The dataset of amazon which contains 21000 reviews out of which 11500 reviews are positive, and 11500 reviews are negative.

Step 3: Data preprocessing

1. Importing the required Libraries

You must download this data set: Data.csv in order to continue

Every time system makes a new model, the system will require to import Numpy and Pandas. Numpy is a library that contains Mathematical functions and is used for scientific computing while Pandas are used to import and manage the data sets.

```
import pandas as pd
import numpy as np
```

Here the system importing the pandas and Numpy library and assigning a shortcut “pd” and “np” respectively.

2. Importing the dataset:

You can easily import an Excel document into Python for the usage of pandas. In order to perform this goal, you’ll want to use examine excel.

before system start, here is a template that you can use in Python to import your Excel report:

```
import pandas as pd
df = pd.Read_excel (r'Path wherein the Excel report is stored file name.Xlsx')
print (df)
```

3. Encoding Categorical Data.

Use values like “Male” and “Female” in mathematical equations of the model so the system needs to encode these variables into numbers. To do this system import the “LabelEncoder” class from the “sklearn preprocessing” library and create an object labelencoder_X of the Label Encoder class. After that system uses the fit_transform method on the categorical features.

4. Splitting the Data set into the Training set and Test Set

Now the system divides our data into two sets, one for training our model called the training set and the other for testing the performance of our model called the test set. The split is generally 80/20. To do this system import the “train_test_split” method of “sklearn. Model_selection” library.

```
from sklearn. model_selection
import train_test_split
```

Now to build our training and test sets, the system will create 4 sets —

1. X_train (training part of the matrix of features),
2. X_test (test part of the matrix of features),
3. Y_train (training part of the dependent variables associated with the X train sets, and therefore also the same indices)
4. Y_test (test part of the dependent variables associated with the X test sets, and therefore also the same indices).

The system will assign to them the test_train_split, which takes the parameters — arrays (X and Y), test_size (Specifies the ratio in which to split the data set).

5. Feature Scaling

Most of the machine learning algorithms use the Euclidean distance between two data points in their computations. As a result, high magnitude characteristics weigh more than low-magnitude characteristics in distance calculations. Standardization or standard Z-score is used to avoid this feature.

This is done through the use of "sklearn. preprocessing" class "StandardScaler."

```
from sklearn.preprocessing import
StandardScaler sc_X = StandardScaler()
```

The further system will transform our X_test set while the system will need to fit as well as transform our X_train set. The transform function will transform all the data to the same standardized scale.

```
X_train = sc_X.fit_transform(X_train)
X_test = sc_X.transform(X_test)
```

Step 4: Feature selection:

Feature extraction is a dimension reduction system through which initial records are reduced to more manageable processing companies. A large number of variables, requiring many computer resources for technical purposes, are a function of these huge data units. The selection of functions means reducing the number of resources needed to describe a wide range of information. One of the main challenges is the number of variables involved when analyzing complex data.

Analysis with a large number of variables generally needs a large amount of storage and computing power, and also a classification algorithm can cause samples to be over-fit training and generate samples poorly. Functional extraction is a general term for methods for building variables to solve these problems while still describing the data accurately. Many practitioners believe that properly optimized functional extraction is the key to efficient model design.

Step 5: Attribute selection:

Removal of the poorly described attributes can significantly improve the accuracy of classification to maintain greater accuracy, as not all attributes are relevant to the classification work and the performance of the used analysis algorithms can decrease irrelevant attributes; for training the classifier an attribute selection scheme was used in this selection.

Step 6: Detection process

Analysis of review through the sentiment analysis using ML techniques. The two supervised machine learning algorithms are SVM and DT-j48 are applied to the dataset. The given review is compared to all the Review_Text in the dataset. Already use the sklearn.module_selection to split the data into trained data and test data. Easy to find the review is fake or original because text converts as 0 and 1. Dictionaries concept is used to assign them and compare the column of Verified_Purchase in the dataset.

Step 7: Comparisons of Result:

Amazon review dataset with different classification Algorithms and the most important classification identified Fake or original algorithm for review detection. [4] To predict which is better to compare the accuracy score of two algorithms.

IV. RESULT ANALYSIS:

The system uses two different supervised machine learning approaches in this section. The text is converted into binary code.

To detect that the review is fake or original, binary code is compared with columns in the dataset. Gender is a data column compared with the review. The result will also show the accuracy score for algorithms and gender. Coding for accuracy score of algorithms

```
In [165]: y_pred[0:5]
Out[165]: array([1, 1, 1, 0, 0], dtype=int64)

In [166]: from sklearn import metrics
          metrics.accuracy_score(y_pred,y_test)*100
Out[166]: 54.885714285714285

In [169]: from sklearn import svm

          model_two = svm.SVC()

          model_two.fit(X_train,y_train)
Out[169]: SVC(C=1.0, break_ties=False, cache_size=200, class_weight=None, coef0=0.0,
              decision_function_shape='ovr', degree=3, gamma='scale', kernel='rbf',
              max_iter=-1, probability=False, random_state=None, shrinking=True,
              tol=0.001, verbose=False)

In [170]: y_pred_two = model_two.predict(X_test)
          metrics.accuracy_score(y_pred_two,y_test)
Out[170]: 0.6428952380952381
```

Fig 4. *Algorithm*

The two supervised machine learning algorithms are SVM and DT-j48 are applied to the dataset. The result of the accuracy of algorithms is shown in the diagram above. The SVM is the best algorithm to decide in that case than the Decision tree. Analysis of review through the sentiment analysis using ML techniques.

a. Build the excel sheet as dataset

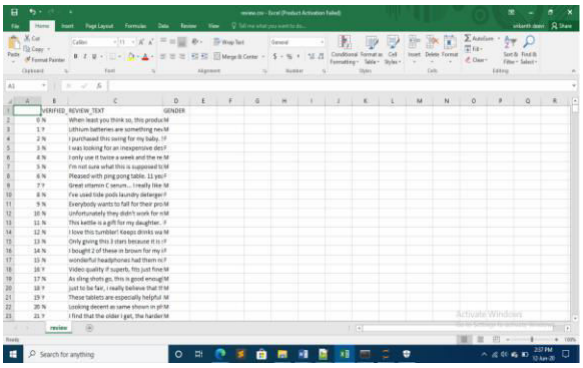

Fig 5*. Dataset*

In the Excel sheet build the amazon dataset. It includes the Verified Purchase columns, review text, Gender as shown in the figure above. Verified purchase is used for determining whether the review is fake or original. Review text is translated to binary code in the dataset. A binary code represents text, instructions to the computer processor, or any other data that uses a two-symbol system. The two-symbol system used on the binary number system is always "0" and "1." The binary code assigns every character, instruction, etc. a binary number pattern, also referred to as bits. Through the dictionary concept, that binary code is compared to gender and verified purchasing columns within the dataset.

b. Graphical user interface

Fig 6*. Result*

The given review is compared to all the Review_Text in the dataset. Already use the sklearn.module_selection to split the data into trained data and test data. Easy to find the review is fake or original because text converts as 0 and 1. The dictionaries concept is used to assign them and compare the column of Verified_Purchase in the dataset. Users search the review that will be compared to predict which one is better using accuracy scores for both algorithms.

V. CONCLUSION AND FEATURE WORK

Included in online shopping, this project is used to identify whether the review is fake or original. Customers write the review in order to identify the gender for future use. The text extraction function is achieved by Tfidvectorizer. To import the label encoder fits the transform review text with preprocessing. To train and check results, using the model selection dataset divided in. The two supervised algorithms for machine learning are applied to datasets. They are Support Vector Machine and a Decision Tree. The metrics module is used to find the score for precision. To show the result build the user interface. User search for the review which matches the training dataset and Displayed in the dataset as original or fake reviews based on verified purchase column. The gender of customers is also displayed in GUI.

For future work, the System would like to expand this project by using other datasets, such as the Flipkart dataset or eBay dataset, and by using different methods of selection. Additionally, the system may use sentiment classification algorithms to detect fake reviews using various tools such as Python and R or R Studio, Statistical Analysis System (SAS), and Stata; fake reviews are also displayed for future use in the output screen. Then the program will use some of these tools to assess the efficiency of our work.

REFERENCE

[1] B. Liu, “Sentiment analysis and opinion mining,” Synthesis lectures on human language technologies, vol. 5, no. 1, 2012, pp. 1–167.

[2] G. Vinodhini and R. Chandrasekaran, “Sentiment analysis and opinion mining: a survey,” International Journal, vol. 2, no. 6, 2012, pp. 282–29.

[3] C. Cortes and V. Vapnik, “SupportVector Networks,” Machine Learning, vol. 20, 1995.

[4] R. Moraes, J. F. Valiati, and W. P. G. Neto, “Document-degree sentiment classification: An empirical comparison between SVM and ann,” ExpertSystems with Applications, vol. 40, no. 2, 2013, pp. 621–633.

[5] P. Kalaivani and K. L. Shunmuganathan, "Sentiment classification of movie reviews by supervised machine learning approaches," Indian Journal of Computer Science and Engineering, vol. 4, no. 4, pp. 285292, 2013.

[6] B. Pang and L. Lee, “A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts,” in Proceedings of the 42nd annual meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2004, p. 271. [Online]. Available from: http://www.cs.cornell.edu/People/pabo/moie-review-data

[7] S. Hassan, M. Rafi, and M. S. Shaikh, "Comparing SVM and naïve Bayes classifiers for text categorization with histology as knowledge enrichment," in Multitopic Conference (INMIC), 2011 IEEE 14th International.IEEE, 2011, pp. 31–34.

[8] C.-H. Chu, C.-A. Wang, Y.-C. Chang, Y.-W. Wu, Y.-L. Hsieh, and W.-L. Hsu, "Sentiment analysis on Chinese movie review with distributed keyword vector representation," in Technologies and Applications of Artificial Intelligence (TAAI), 2016 Conference on.IEEE, 2016, pp.84–89.

[9] V. Singh, R. Piryani, A. Uddin, and P. Waila, "Sentiment analysis of the movie reviews and blog posts," in Advance Computing Conference(IACC), 2013 IEEE 3rd International. IEEE, 2013, pp.

893–898.

[10] G. Vinodhini and R. Chandrasekaran, "Sentiment analysis and opinion mining: a survey," International Journal, vol. 2, no. 6, 2012, pp. 282–292. [14] G. Xu, Y. Cao, Y. Zhang, G. Zhang, X. Li, and Z. Feng, "Trm:

Computingreputation score by mining reviews." in AAAI Workshop: Incentives and Trust in Electronic Communities, 2016.

[11] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H.Witten, "The weka data mining software: an update," ACM SIGKDDexplorations newsletter, vol. 11, no. 1, 2009, pp. 10–18.

[12] A. Abdel-Hafez and Y. Xu, "A survey of user modeling in social media websites," Computer and Information Science, vol. 6, no. 4, 2013, p. 59.

[13]Evangelos Giannitsis [1], Matthias Mueller-Hennessen [2], Tanja Zeller "Gender-specific Reference Values for High-Sensitivity Cardiac Troponin T and I in Well-Phenotyped Healthy Individuals and Validity of High-Sensitivity Assay Designation" 10.1016/j.clinbiochem. 2019.11.013. Epub 2019 Nov 28.

[14] Hiromasa Takahashi, Junyi Shen & Kazuhito Ogawa "Gender-specific reference-dependent preferences in the experimental trust game" in Published on10 January 2020