Review spam detection using machine learning

Project report

BACHELOR OF TECHNOLOGY

In

Chemical Engineering

Submitted by

Dipankar Modak

160903112

Under the guidance of

Dr.Krishnamoorthi Makkithaya

Professor

Department of Computer Science

Manipal Institute of Technology, Manipal

DEPARTMENT OF CHEMICAL ENGINEERING
MANIPAL INSTITUTE OF TECHNOLOGY
(A constituent unit of MAHE), MANIPAL-576104

31 May, 2020

SYNOPSIS

With the ever-increasing popularity of review websites that feature user-generated opinions there is a greater chance to do opinion spam also known as review spamming. Opinion spam is self- promotion of Product/Service done by individuals hired by companies to provide a positive feedback to otherwise poorly reviewed product. A customer goes through the online reviews that the available for that product/service and based on that make decisions on whether to purchase the product/service. This dependency has paved the way for fraudsters to gain incentives in exchange to putting up fake information to mislead customers.

Review spam can financially affect businesses and might cause a sense of mistrust in the general public, therefore, due to its significance, this problem has recently attracted the consideration of the media and governments as well. In this project we concentrate on more of an insidious type of opinion spam i.e. deceptive opinion spam that has been written to sound authentic to the reader in order to deceive them. For this project we consider dataset containing reviews of hotels gathered from Trip Advisor and Amazon Mechanical Turk. In this project we use combination of features like Bag of words, Term –Frequency Inverse Document frequency, Review length, Sentiment polarity and Parts of speech counts for each documents to provide more number of features to the learning model.

For this project we use two machine learning model i.e. Naïve Bayes Classifier and Logistic Regression Classifier. Two of them are quite different. Naïve Bayes is a generative learning algorithm while Logistic regression is a discriminative algorithm. We start from Naïve Bayes Classifier and a simple feature extraction technique in Natural language processing called Bag of words and continues to add features it until the accuracy of the model becomes acceptable.

CERTIFICATE

This is to certify that the project work on title "*Review spam detection using machine learning*" submitted to the MAHE during May, 2020 for the partial fulfillment of Bachelor of technology in Chemical Engineering by **Dipankar Modak** is a bonafide record of original work carried out by him during the period from January, 2020 to May, 2020. The work reported herein does not form part of any other thesis or dissertation on the basis of which any degree, diploma, associate ship, fellowship or other titles were awarded.

Dr.Krishnamoorthi Makkithaya

(Signature)

Professor
Department of Computer Science
Manipal Institute of Technology,
MAHE, Manipal

Dr. S.V.S.R Krishna Bhandaru

(Signature)

Head of Department
Department of Chemical Engg.
Manipal Institute of Technology,
MAHE, Manipal

ACKNOWLEDGEMENT

I express my regards and a great sense of gratitude to **Dr. Krishnamoorthi Makkithaya**, Professor, Manipal Institute of Technology for his valuable guidance, moral support, and supervision throughout the course.

I take this opportunity to express gratitude to all of the Department faculty members and staff especially **Dr S.V.S.R Krishna Bhandaru** for his kind help and support.

I am also grateful to my Family for their constant encouragement, guidance and support to help me come out with flying colors in tough times.

(Signature of the student)

LIST OF TABLES

LIST OF FIGURES

Contents

CHAPTER 1
INTRODUCTION

1. USER REVIEW

A user review is a review conducted by a consumer and published to a review site following buying a product or the evaluation of a service. As, more and more individuals and organizations have become accustomed to consulting user generated reviews before making purchases or online bookings. Considering great commercial benefits, merchants, however, have tried to hire people to write undeserving positive reviews to advertise their products or services, and meanwhile post malicious negative reviews to defame those of their competitors. Those spam reviews and opinions, which are deliberately produced in order to promote or demote targeted product or services, are known as **deceptive opinion spam.**

2. TYPES OF SPAM REVIEWS

There are 3 types of Opinion spam reviews-(1) Fake reviews, (2) Brand reviews,(3) Non – reviews.

(1) Fake reviews-In this type of reviews the user doesn't have the experience of the product/services that they are writing about. There is an insidious agenda behind it, either to promote a product or defame it.

(2) Brand reviews- This user reviews are solely written by user on the basis of prior experience with the product/services of the particular brand and has nothing to do with that particular product. Past experience is the key here.

(3) Non reviews-This are the user reviews that doesn't contains any relevant information or sentiment with respect to either the brand or the product. They mostly consist of advertisement of their product or services.

The first type of reviews are the most difficult to detect. Fake reviews are the worst type of advertisement as they directly impact the reputation of a product/service. Whereas the (2) and (3) type of reviews are quite rare and have hardly any effect on the sentiment of the customers.

3. MOTIVATION

Hotels or restaurants are vulnerable to fake reviews, particularly if they are negative. While rare, it can happen, whether it's an unethical competitor or an individual who has decided to cause problems for a business. A study conducted by Bright Local research revealed 82% of consumers read a fake review. Among youngsters the distribution was even higher. Allowing customers to be exposed to an increasing number of fake reviews, it's perfectly highly likely that we'll soon begin to see trustfulness in peer-to-peer recommendations being eroded. Therefore they should have a detection algorithm which filters out the spam reviews from the genuine reviews and improve the customer services.

4. OBJECTIVE OF THE PROJECT

1. Extraction of text features from the review text.
2. Application of suitable machine learning algorithm on the extracted features and increase the accuracy of the classification models.

CHAPTER 2
BACKGROUND THEORY

1. LITERATURE REVIEW

The first method for review spam detection was proposed in 2007 by Jindal and Liu .This paper was followed by [1] and [2] in which initial ideas were further investigated. Jindal et al. demonstrated that on data collection of 5.18 million reviews and 2.14 million reviewers from amazon (a e-commerce platform) showed duplication of reviews by using the method of spam detection. [2]. According to Dixit et al. he proposed that spam review can be divided into three cluster, (1) Reviews on Brands – failure of reviewing of the product as the reviews were only directed towards the seller or the brand, (2) Deceptive Reviews – was one of the main focus of this paper, and (3) Non-Reviews – comments that were majorly either irrelevant to the product or advertisements. [Dixit S, Agrawal AJ. Survey on review spam detection. Int J Comput Commun Technol ISSN (PRINT). 2013 Jun;4:0975-7449.]: Fei et al. [3]observed that using only word features like Bag of words(BOW) alone proved inadequate for machine learning algorithms when learners were trained using synthetic fake reviews, since the features being created were not present in real-world fake reviews. Ott et al. [4] Tausczik et al. employed psycholinguistic features on the basis of features generated by LIWC [5]combined with standard word and Part of Speech (POS) n- gram features. Mukherjee et al. [6]extend that work including also style and POS based features, such as deep syntax and POS sequence patterns. The techniques that were supervised like Liu et al.[7]used a Bayesian approach and laid out a clustering problem with opinion spam sensing. Li et al. [4]

2. SYNOPSIS OF LITERATURE REVIEW

From the Literature survey, it was understood on which dataset the project would be worked on. Reviews have been broadly classified in 3 categories- 1.Untruthful Reviews, 2.Non reviews, and 3.Reviews on brand. All of papers focused on Review centric features like Bag of Words or Parts of speech and found that the accuracy of the model with these features alone were quite poor. Therefore they tried to implement other stylometric features like length of words and sentences. Additional features such as such as Maximum content similarity can be employed provided the dataset contains information regarding the reviewer. Naive Bayes being the choice of algorithm they wanted to implement in supervised dataset.

CHAPTER 3
METHODOLOGY

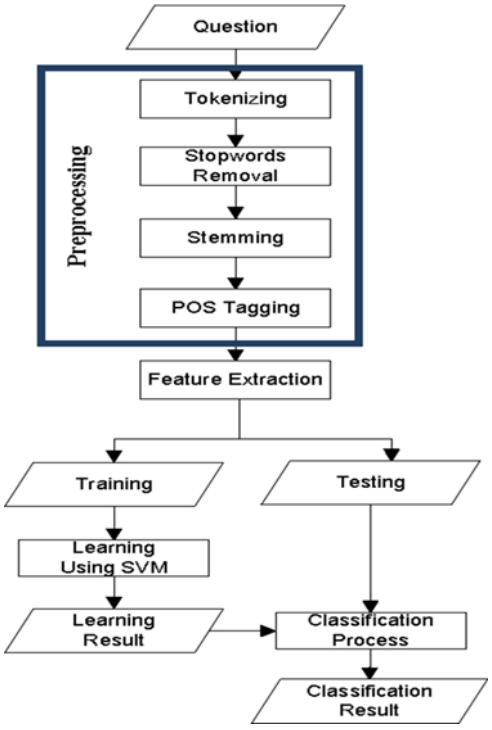

Fig-3.1 Flowchart of the Methodology to be followed.

1. DESCRIPTION OF THE FLOWCHART

The Fig 3.1 represent the process in which the the classification model training and testing is done.At first the review text are extracted and then preprocessed to remove all the unwanted or non essential words,punctuations and other operations.Next step is feature extraction where the textual features are converted to vectorial representation.Then the generated matrix is then splitted to 2 parts i.e training and testing set.Model building is done on the training set with a choosen Machine learning algorithm.The build model is then tested on the testing set whose classification performance can be examined using a confusion matrix.

2. CHOICE OF PROGRAMMING LANGUAGE

Till mid-term all the programming was done on R primarily because it has great visualization tools called ggplot2 and most of the report was based on data visualization .But python was choosen later in the project because of NLTK Toolkit it provides.It is easier to use for someone who is beginner at Natural Language processing to implement machine leaning algorithm on text data. Numpy also provides a great help in representing the data into vectors and matrices.

3. SELECTING THE RIGHT DATASET

Data used in training and testing systems for spam detection comes from myriad of sources. Lack of standard datasets for this type of problem makes training of model under supervised learning algorithm difficult. The dataset used for this project consist of truthful and deceptive opinion reviews from 20 Chicago hotels collected from AMT and Yelp. It has been obtained from https://www.kaggle.com/rtatman/deceptive-opinion-spam-corpus

3.1 Reading the dataset.

This dataset contains:

1. 400 truthful positive reviews from Trip Advisor
2. 400 deceptive positive reviews from Mechanical Turk.
3. 400 truthful negative reviews from Expedia, Hotels.com, Priceline, Trip Advisor and Yelp
4. 400 deceptive negative reviews from Mechanical Turk.

Attributes of the dataset are-

1. *Label*- Describes if the Review text is classified as “truthful” or “deceptive”.
2. *Hotel*- Describes the Name of the Hotel for which the Review was written.
3. *Polarity*- Describes the tone of the Review Text-“Positive” or “Negative”
4. *Source*- Mentions the website from which the Review text was extracted.
5. *Review text*- A textual representation of reviewer sentiment, emotions.

The dimension of the dataset is -1600 rows and 5 columns.

Index	Label	hotel	polarity	source	Reviewtext
0	truthful	conrad	positive	TripAdvisor	We stayed for a one night getaway w
1	truthful	hyatt	positive	TripAdvisor	Triple A rate with upgrade to view
2	truthful	hyatt	positive	TripAdvisor	This comes a little late as I'm fir
3	truthful	omni	positive	TripAdvisor	The Omni Chicago really delivers or
4	truthful	hyatt	positive	TripAdvisor	I asked for a high floor away from
5	truthful	omni	positive	TripAdvisor	I stayed at the Omni for one night
6	truthful	conrad	positive	TripAdvisor	We stayed in the Conrad for 4 night
7	truthful	omni	positive	TripAdvisor	Just got back from 2 days up in Chi
8	truthful	omni	positive	TripAdvisor	We arrived at the Omni on 2nd Septe

Fig-3.2. Screenshot of the dataset

3.2 Pre-processing of the dataset.

Following are the pre-processing steps that are done-

1. *Removal of Stop words*- Stop words are frequent words in any language which may not be useful or bring any valuable information in text .eg-“a”,”an”,”the”.

2. *Stemming*- Stemming is the process of reducing words to their word stem; base or root, form generally a written word form e.g.- Consulting &Consultant have the same root word Consult.

3. *Tokenization*- Tokenization is the process of splitting a string, text into a list of words.

4. *Parts of speech tagging (POS)*- POS tags are labels given to words on basis of their context. This process includes tagging a word based on its definition and relationships with adjacent words. Words then are marked as Noun, Pronouns, etc. This information is collected and with additional features fed into a machine learning algorithm.

5. *Removal of Numbers and Punctuation*

4. FEATURE EXTRACTION

Features that can be used to spam detection based on review content are as follows.
Review centric features-
4.1 Bag of Words (BOW):is a features extraction technique that is used for training machine learning algorithms. It creates vocabulary of all the distinct words occurring in all the reviews in the training set. Each unique word in the corpus is represented as a feature vector based on their occurrences in each of the review.

	0	1	2	3	4	5
0	1	0	0	0	0	0
1	0	0	0	0	0	0
2	0	0	0	0	0	0
3	0	0	0	0	0	0
4	0	0	0	0	0	0
5	0	0	0	0	0	0
6	0	0	0	0	0	0
7	0	0	0	0	0	0

Fig 3.3 Bag of words model

Here each row index represent a particular review and each column consist of a unique word represent as number. 1 represent that a particular word occurred once in that particular review. 0 means absence.

4.2 Term frequency Inverse Document frequency (TFIDF) is a technique used to help compensate for words found relatively often in different reviews which makes it hard to distinguish between the reviews because they are too common. The Greater the frequency of a word in a review, the more important it is to the review. However the measurement is offset by the review size- the total number of words, the review contains- and by how often the word appears in other reviews. Similar to BOW each row index represents a review and each column represents a particular word.

	1443	1444	1445	1446	1447	1448
0	0	0	0	0	0	0
1	0	0	0	0	0	0
2	0	0	0	0	0	0
3	0	0.245893	0	0	0	0
4	0	0	0	0	0	0
5	0	0	0	0	0	0
6	0	0	0	0	0	0
7	0	0	0.181523	0	0	0

Fig 3.4 Term frequency Inverse Document Frequency Model

4.3 Parts of speech tags counts (POS) Here each row index represents a particular review and each column represents a particular part of speech in English language. The matrix has dimension of (1600*31) where 1600 represent review counts and 31 represents 31 parts of speech in English Language. Each cell gives frequency of occurrence of a particular part of speech.

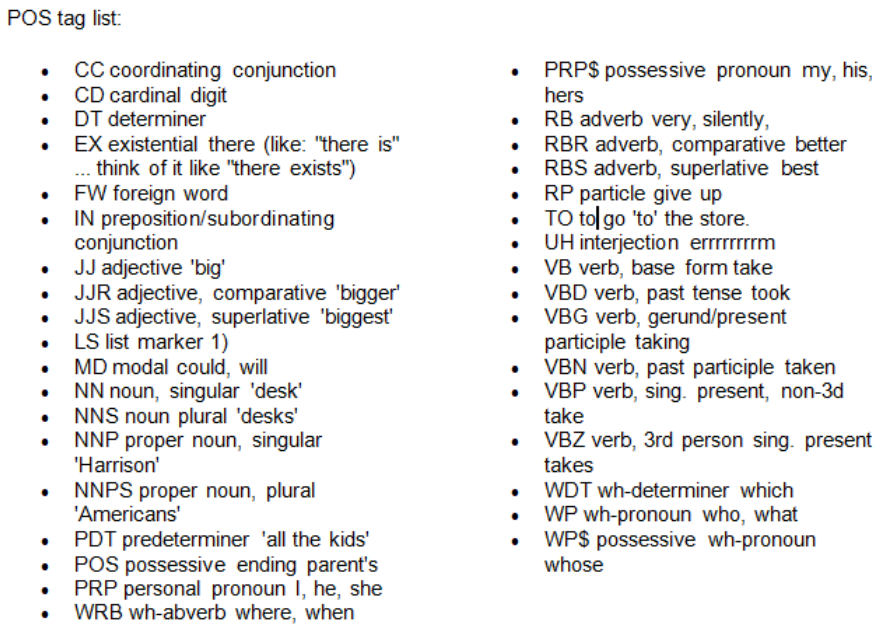

Fig 3.5 Parts of speech tag list in English

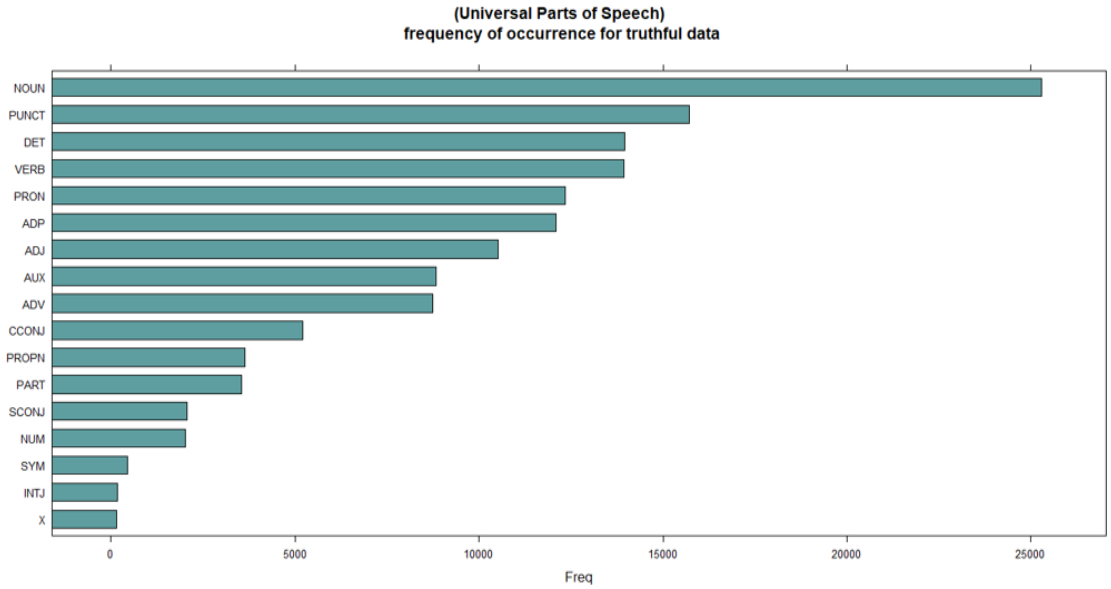

Fig 3.6 Parts of speech frequency in Truthful data

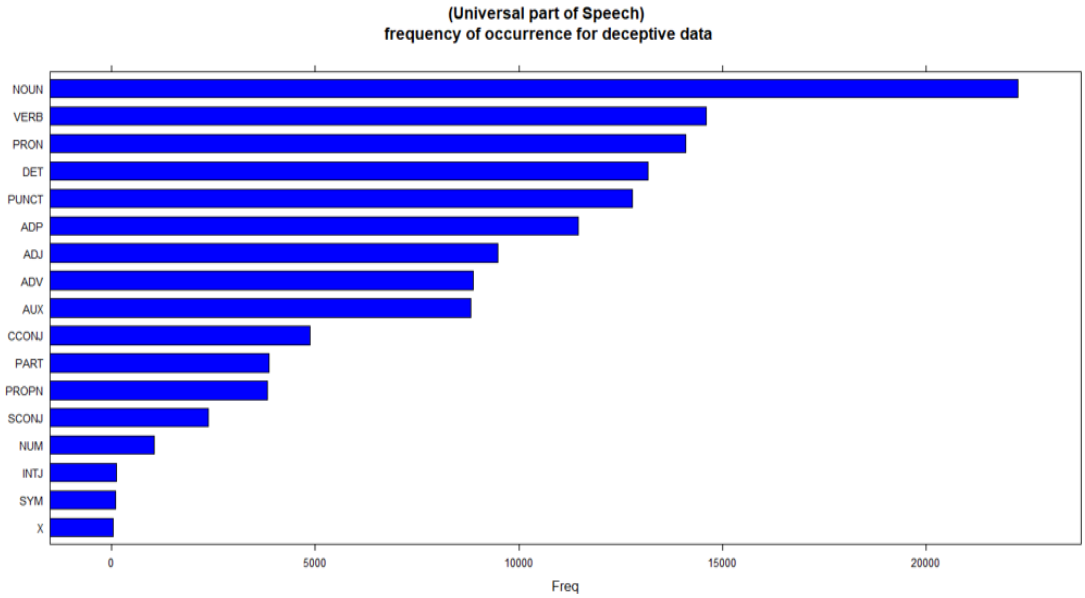

Fig 3.7 Parts of speech frequency in Deceptive data

	0	1	2	3	4	5
0	0	4	0	0	0	2
1	0	0	0	0	0	1
2	0	1	0	0	1	5
3	0	1	0	0	0	0
4	0	1	0	0	0	0
5	0	3	1	0	0	3
6	0	0	0	0	0	0
7	0	2	0	0	1	2

Fig 3.8 Parts of Speech counts model.

4.4 Review length (RL)-

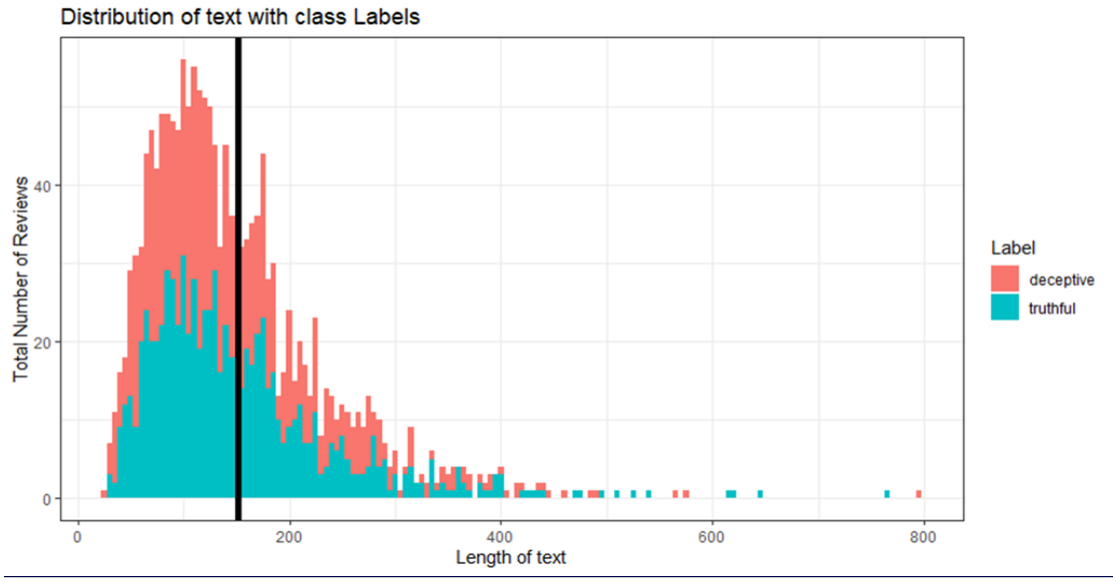

Fig 3.9 Review length vs. Class Label

According to recent studies ,80% of spammers have reviews no longer than 135 words while more than 92% of reliable reviewers have review length of greater than 200 words. The above graph represents the frequency distribution of our dataset. As we can observe that lesser the review length more the probability of the review being deceptive and higher the review length the probability of the review being deceptive drastically decreases. This can be used as a feature.

4.5 Sentiment Polarity (SP) in NLP is calculated by difference between numbers of positive words and number of negative words. The range of values lies between -1 to 1, where -1 represent a very negative sentiment while 1 represent an extremely positive sentiment. Neutrality is represented by a score of 0. Any score on the extremity is highly unlikely to be a genuine review. These features can be used to detect the brand reviews where the customer writes extremely positive reviews solely on the basis of brand name. The sentiment score can be calculated by other method but for this project we kept the calculation as simple as possible by cumulative summation of both positive and negative reviews.

4.6 Edit Distance(ED) is a way of quantifying how dissimilar two strings (e.g. words) are to one another by adding the minimum number of operations required to convert one string into the other. Basically it is minimum number of changes a given string has to be done to convert into a target string. These changes can be implemented by 3 operations- Insert, Remove and Replace. Each of the operations is of same cost.

	0	1	2	3	4	5
0	0	54	97	62	56	86
1	54	0	102	61	32	89
2	97	102	0	101	102	101
3	62	61	101	0	59	86
4	56	32	102	59	0	89
5	86	89	101	86	89	0
6	56	58	99	61	58	86
7	67	65	99	67	66	88

Fig 3.10 Minimum Edit Distance

Here each row and column index represents a review and their corresponding value represents the edit distance between them. The higher the edit distance the higher is the dissimilarity between two reviews. So minimum edit distance has been taken for each reviews and used as feature vector of our model.

```
In [83]: hamsum=np.sum(dataham,axis=1)
    ...: np.sum(hamsum)
Out[83]: 113590758
```

```
In [85]: spamsum=np.sum(dataspam,axis=1)
    ...: np.sum(spamsum)
Out[85]: 117261780
```

Fig 3.11 Total minimum edit distances in ham and spam data

5. SPLITTING THE DATA

Set Division Guidelines for Prediction Study Design

- For large sample sizes
  - 60% training
  - 20% test
  - 20% validation
- For medium sample sizes (Our dataset falls in these range)
  - 60% training
  - 40% test

6. ALGORITHM SELECTION

6.1 Naïve Bayes

Goal in selecting models is to avoid over fitting on training data and minimize error on test data. From, Literature survey, there are two algorithms which provided the highest accuracy for spam detection on this dataset. Naive-Bayes and later Logistic Regression .For Naïve Bayes the Mathematical equation is represented by-

$$\pi_k P(X_1, \ldots, X_m \mid Y = k) = \pi_k P(X_1 \mid Y = k) P(X_2, \ldots, X_m \mid X_1, Y = k)$$
$$= \pi_k P(X_1 \mid Y = k) P(X_2 \mid X_1, Y = k) P(X_3, \ldots, X_m \mid X_1, X_2, Y = k)$$
$$= \pi_k P(X_1 \mid Y = k) P(X_2 \mid X_1, Y = k) \ldots P(X_m \mid X_1 \ldots, X_{m-1}, Y = k)$$

for predictors *X*1. . .*Xm*, we want to model *P*(*Y* = *k* | *X*1, . . ,*Xm*) however, if we make the assumption that all predictor variables are independent to each other, the quantity can be simplified to-

$$\pi_k P(X_1, \ldots, X_m \mid Y = k) \approx \pi_k P(X_1 \mid Y = k) P(X_2 \mid Y = k) \ldots P(X_m \mid , Y = k)$$

Naïve Bayes function

P(*Y* = *k*), is determined from the data to be some known quantity Π*k* (also known as prior probability).

6.2 Logistic Regression

Algorithm is similar to linear regression, with the only difference being the y data which should contain integer values and represents class. Here the two classes are separate by a decision boundary (a linear line).It consist of sigmoid function and an error function. Less the value of cost function more accurate will be the model. Regularization parameter can be used to reduce over fitting on the training data.

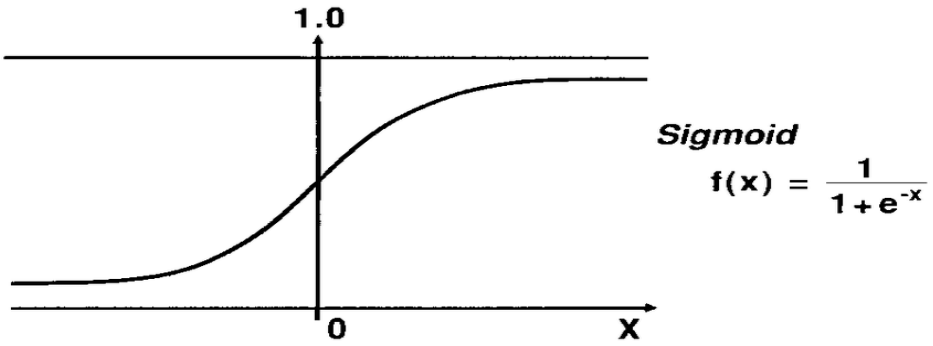

Fig 3.12 Sigmoid Function

$$\text{Cost}(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$

Fig 3.13 Cost Function of Logistic Regression

7. DIMENSION REDUCTION

Principal Component analysis (PCA)-Constructing a classification model may not require every feature. Ideally we want to capture the most variation in data with the least number of variables.PCA is suited to do this and will help reduce number of features as well as reduce noise

Linear discriminate analysis (LDA)-is a dimension reduction technique where a new feature space is found out in order to maximize the class separability in the data. Specifically, the model wants to find a linear combination of input features that achieves the maximum separation for databetween classes and the minimum separation of samples within each class.

8. TRAINING THE MODEL

Here we use sckitlearn package in Python which consist of many data pre-processing steps and machine learning model. Link for sckitlearn information page-https://scikit-learn.org/stable/

9. PERFORMANCE MEASURES

1. *Accuracy*-is the ratio of number of correct prediction to total number of predictions. As we increase the number of features, accuracy also tends to increase. This may cause overfitting of our model which would provide poor performance result on test data. Therefore along with accuracy sensitivity and specificity should also be high for optimal model performance.

2. *Sensitivity*-is the ratio of items predicted as positives which are actually positive versus total number of positive items.

3. *Specificity*- is the ratio of items predicted as negative which are actually negative versus total number of negative items.

CHAPTER 4
RESULTS & DISCUSSION

1. WORKDONE

- Testing classifier on the test set.
- Combining individual features to generate a large feature vector
- Analysis accuracy of models using different combination of features
- Selecting the best classifier and best combination of features for spam detection

2. ANALYSIS OF RESULTS

2.1 Accuracy of different models

2.1.1 Naïve Bayes

1. *BOW*–Matrix that contains mostly zero values are called sparse matrix. Features like bag of words generate a huge sparse matrix. So computation becomes difficult and time expensive. Also it reduces the accuracy of the model as observed from the graph from our data. So to increase the accuracy we reduce the matrix with sparsity of 0.9732 (min df =5).Therefore obtain an accuracy of 0.692.

Table 4.1 Performance of Bow model

Accuracy	0.692
Sensitivity	0.613
Specificity	0.765

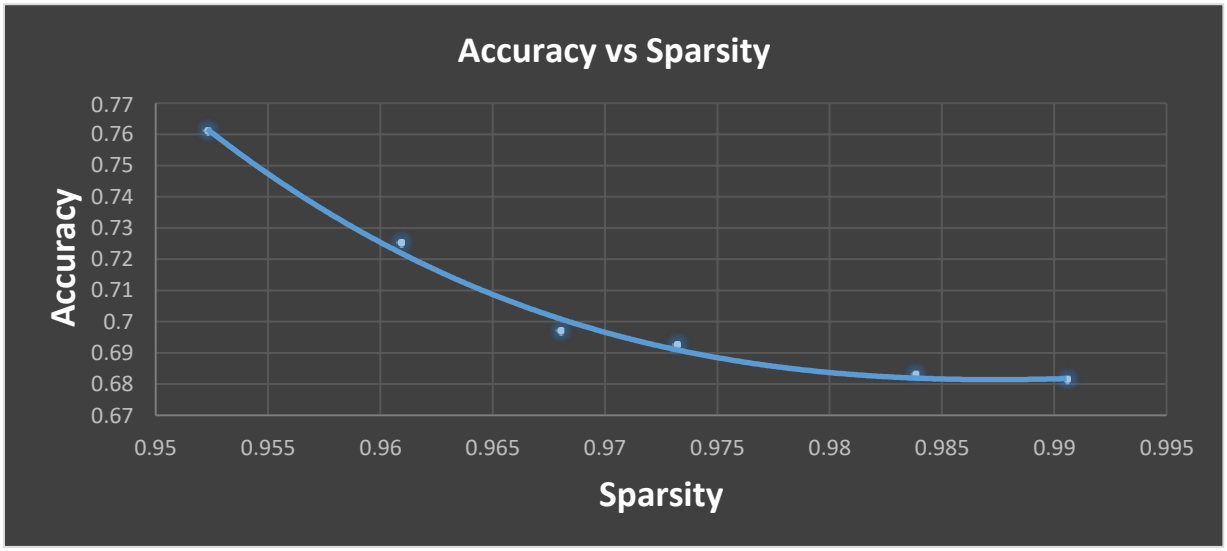

Fig 4.1 Accuracy vs. Sparsity

2. *TFIDF+BOW*-By combining both the features the Bag of words and TFIDF ,now each of the words are represented by their weights in a particular review in addition to the occurrences. Accuracy increased to 0.693.

Table 4.2 Performance of Bow+TFIDF model

Accuracy	0.693
Sensitivity	0.659
Specificity	0.725

3. *POS+TFIDF+BOW*-Parts of speech counts have been added to the feature vector which initially consist of TFIDF+BOW features. Accuracy improved to 0.695.

Table 4.3 Performance of Bow+TFIDF+POS model

Accuracy	0.695
Sensitivity	0.717
Specificity	0.674

4. *POS+TFIDF+BOW+RL*-With the review length being added the dimensionality of the feature vector increased. Till now the feature vector primarily consist vector input from-1 to 1, but review length has very different set of input from-1 to 1. Since we are using Naïve Bayes algorithm, we can skip feature scaling as it is a graphical model based classifier. Accuracy improved significantly to 0.747.

Table 4.4 Performance of Bow+TFIDF+POS+RL model

Accuracy	0.747
Sensitivity	0.711
Specificity	0.744

5. *POS+TFIDF+BOW+RL+SP*-The sentiment function in NLTK Toolkit will return a named tuple of form Sentiment (polarity, subjectivity). The polarity score is a float number within the range (-1.0, 1.0). we consider only the polarity score. Accuracy remains the same. So this feature might not be helpful and acts as noise.

Table 4.5 Performance of Bow+TFID+POS+RL+SP model

Accuracy	0.747
Sensitivity	0.711
Specificity	0.744

6. *POS+TFIDF+BOW+RL+SP+ED*- Edit distance is new feature that is added on to the the feature vectors. Accuracy of the model is increased to 0.753.

Table 4.6 Performance of Bow+TFIDF+POS+RL+SP+ED model

Accuracy	0.753
Sensitivity	0.668
Specificity	0.834

7. *Model selection and Feature Combination*

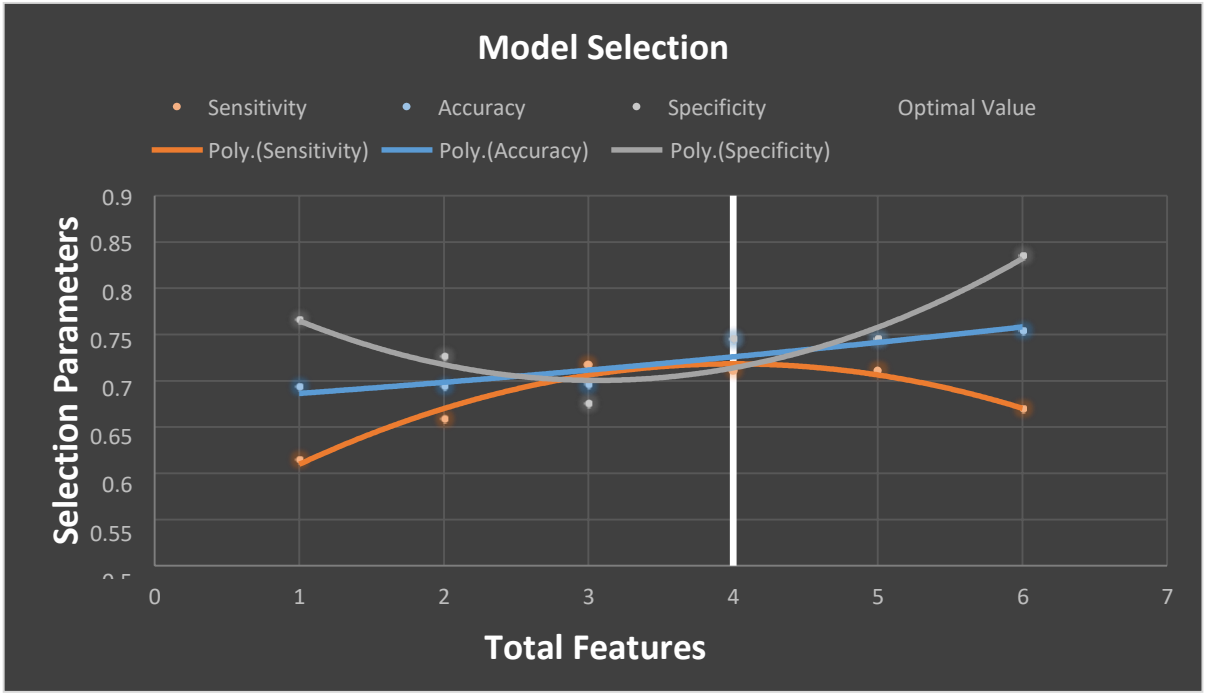

Fig 4.2 *Model* Selection (Naïve Bayes)

Here we observe that accuracy of our model continues to increase as the number of features in our model increases. But sensitivity of our model first increases and then reduces. Specificity first decreases then increases. Therefore the best model is the one which have a higher accuracy with reduced cases of false positives and false negatives i.e point of intersection of the three curves. In this case we observe from the plot that when number of features we used is four, the model performance is overall the best. Therefore for Naïve Bayes classifier a feature combination of (TFIDF+BOW+POS+RL) generates a good model.

2.1.2 Logistic regression

For the sake of visualizing the decision boundary, we convert our large feature vector to 2 dimensions feature vector using PCA. Then the training set was used to build the model using Logistic Regression algorithm and then test on the testing set. Following is the plot of the decision boundary of our dataset. Here PC1 represents the first feature vector and PC2 represents the second feature vector. The separation boundary between red and green region is the linear decision boundary. 0 and 1 are the two classes i.e. Truthful and Deceptive.

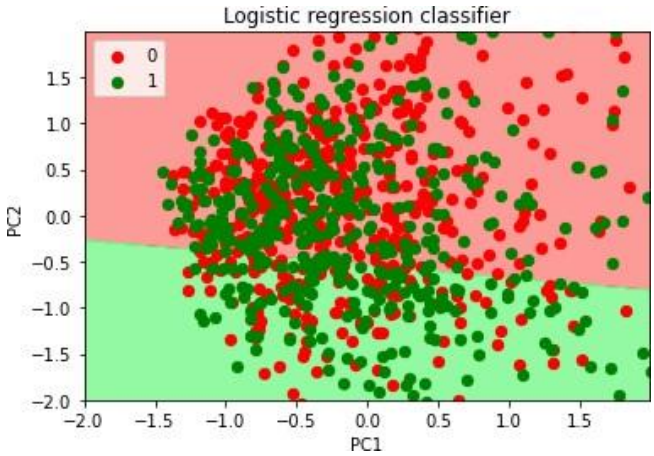

Fig 4.3 2D plot of decision boundary by Logistic Regression

1. *BOW*- Here the regularization parameter is set to one with number of iterations to be 100.Sparsity is 0.97. Accuracy is about 0.856.

Table 4.7 Performance of Bow model

Accuracy	0.856
Sensitivity	0.827
Specificity	0.882

2. *TFIDF+BOW*-Same as before with regularization parameter set to one and number of iterations set to 100.Accuracy increases to 0.876.

Table 4.8 Performance of TFIDF+Bow model

Accuracy	0.876
Sensitivity	0.879
Specificity	0.873

3. *TFIDF+BOW+POS*-Here the accuracy of model decreases in contrast to Naïve Bayes model. Therefore POS acts as a noise and is not helpful for classification. Accuracy is 0.835

Table 4.9 Performance of TFIDF+BOW+POS model

Accuracy	0.835
Sensitivity	0.818
Specificity	0.852

4. *TFIDF+BOW+POS+RL*-Accuracy increases to 0.842

Table 4.10 Performance of TFIDF+BOW+POS+RL model

Accuracy	0.842
Sensitivity	0.824
Specificity	0.858

5. *TFIDF+BOW+POS+RL+SP*-Same as Naïve Bayes, SP acts as a noise for both the classifier and therefore a bad feature vector altogether. Accuracy is 0.842.

Table 4.11 Performance of Bow +TFIDF+POS+RL+SP model

Accuracy	0.842
Sensitivity	0.824
Specificity	0.858

6. *TFIDF+BOW+POS+RL+SP+ED*-Number of iterations have been increased to 1000 to reach the convergence of the cost function. Accuracy has increased to 0.846

Table 4.12 Performance of Bow+TFIDF+POS+RL+SP+ED model

Accuracy	0.846
Sensitivity	0.84
Specificity	0.852

7. *Model selection and feature combination*

Fig 4.4 Model selection (Logistic Regression)

Here we observe that accuracy of our model first increases then decreases as the number of features in our model increases. Sensitivity of our model also first increases then decreases with model complexity. Specificity declines with model complexity. Therefore the best model is the one which have a higher accuracy and also has high sensitivity and specificity. In this case we observe from the plot that when number of features we used was two, the model performance is overall the best. Therefore for Logistic Regression algorithm a feature combination of (TFIDF+BOW) generates a good model.

3. COMPARISON OF LOGISTIC REGRESSION AND NAIVE BAYES

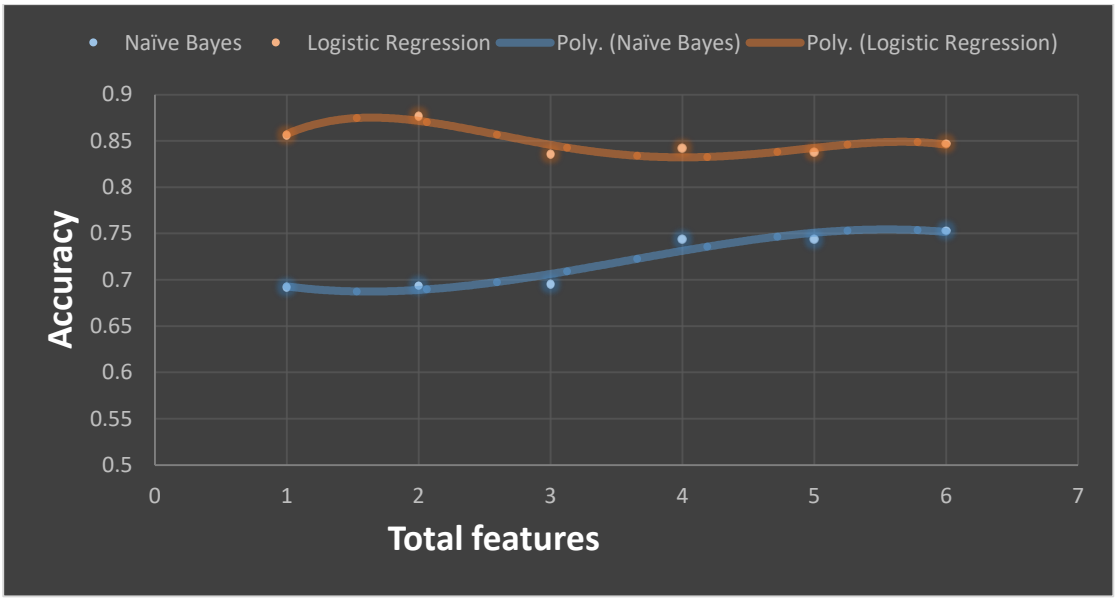

Fig 4.5 Comparison of Naïve Bayes and Logistic Regression

Logisitic Regression is a discrimnative algorithm meaning it models the decision boundary between the classes. Whereas Naïve bayes is a generative algorithm which model the distribution of each class.Since our dataset had huge number of features from the very first model, Logistic Regression performed better than Naïve Bayes as discriminative algorithm performs better as the training size increases.Therefore more likely to overfit the training data and, hence higher accuracy compared to Naïve Bayes.

4. REDUCING MODEL COMPLEXITY

The combination of features vectors of TFIDF+BOW+POS+RL+SP+ED generates a large matrix(1600*3719). Applying LDA to the combined feature vector the dimension reduces to (1600*1).LDA select the best feature which can separate the two classes of Label.Accuracy score improves drastically when extracted feature is applied to both Naïve Bayes and Logistic Regression to 0.987. Also specificity and sentitvity are both very high.

Table 4.13 Accuracy of model after LDA

Accuracy	0.987
Sensitivity	0.993
Specificity	0.981

CONCLUSIONS

1. Logistic Regression learns from the training data taking into account all the possible correlations between features. While in Naïve Bayes there is a perquisite assumption that all the features act independently to each other. Therefore Logistic Regression is more prone to overfitting the data than Naïve Bayes. This is the reason why there is higher accuracy in case of the former than the latter.

2. TFIDF+BOW+POS+RL feature combination generates best model with Naïve Bayes algorithm .TFIDF+BOW feature combination generates best model with Logistic Regression.

3. Though accuracy should increase with increase in features, we have seen that adding sentiment polarity on both Naïve Bayes and Logistic regression algorithms reduced the model performance. It is a noise and doesn't provide any meaningful information to our classifier irrespective of the classifier.

4. Dimension reduction such as LDA can be applied to generate a 1D feature vector from a large matrix. This generated feature can later be used for training models with Logistic Regression and Naïve Bayes which increases the accuracy significantly.

REFERENCES

Reference to a journal publication:

[1]N. Jindal and B. Liu, "Analyzing and Detecting Review Spam," in *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, Oct. 2007, pp. 547–552, doi: 10.1109/ICDM.2007.68.

[2]N. Jindal and B. Liu, "Opinion spam and analysis," in *Proceedings of the 2008 International Conference on Web Search and Data Mining*, Palo Alto, California, USA, Feb. 2008, pp. 219–230, doi: 10.1145/1341531.1341560.

[3]G. Fei, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos, and R. Ghosh, "Exploiting Burstiness in Reviews for Review Spammer Detection," in *Seventh International AAAI Conference on Weblogs and Social Media*, Jun. 2013, Accessed: Jun. 02, 2020. [Online]. Available: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6069.

[4]J. Li, M. Ott, C. Cardie, and E. Hovy, "Towards a General Rule for Identifying Deceptive Opinion Spam," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Baltimore, Maryland, Jun. 2014, pp. 1566–1576, doi: 10.3115/v1/P14-1147.

[5]Y. R. Tausczik and J. W. Pennebaker, "The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods," *J. Lang. Soc. Psychol.*, vol. 29, no. 1, pp. 24–54, Mar. 2010, doi: 10.1177/0261927X09351676.

[6]A. Mukherjee, V. Venkataraman, B. Liu, and N. Glance, "What Yelp Fake Review Filter Might Be Doing?," in *Seventh International AAAI Conference on Weblogs and Social Media*, Jun. 2013, Accessed: Jun. 02, 2020. [Online]. Available: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6006.

[7]G. Wang, S. Xie, B. Liu, and P. S. Yu, "Review Graph Based Online Store Review Spammer Detection," in *2011 IEEE 11th International Conference on Data Mining*, Dec. 2011, pp. 1242–1247, doi: 10.1109/ICDM.2011.124.

Reference to a website:

[1]https://towardsdatascience.com/implementing-a-naive-bayes-classifier-for-text-categorization-in-five-steps-f9192cdd54c3

[2] https://towardsdatascience.com/the-naive-bayes-classifier-e92ea9f47523

[3] https://towardsdatascience.com/machine-learning-text-processing-1d5a2d638958

[4] https://www.aclweb.org/anthology/P14-1147.pdf

[5] https://rpubs.com/cen0te/naivebayes-sentimentpolarity

[6]https://stats.stackexchange.com/questions/323530/na%C3%AFve-bayes-theorem-for-multiple-features

PROJECT DETAILS

Student Details			
Student Name	Dipankar Modak		
Register Number	160903112	Section / Roll No	44
Email Address	dipankar.kota@gmail.com	Phone No (M)	6901255886
Student Name			
Register Number		Section / Roll No	
Email Address		Phone No (M)	
Project Details			
Project Title	Review spam detection using machine learning		
Project Duration	4 months	Date of reporting	6/06/2020
Organization Details			
Organization Name	Manipal Institute of Technology		
Full postal address with pin code	Udupi-Karkala Road.Eshwar Nagar ,Manipal,Karnataka,576104		
Website address	https://manipal.edu		
Supervisor Details			
Supervisor Name	Dr.Krishnamoorthi Makkithaya		
Designation	Professor		
Full contact address with pin code	Department of Computer Science, Manipal Institute of Technology,Manipal,576104		
Email address	k.moorthi@manipal.edu	Phone No 0944891148	
Internal Guide Details			
Faculty Name			
Full contact address with pin code	Department of Chemical Engineering, Manipal Institute of Technology, Manipal – 576 104 (Karnataka State), INDIA		
Email address			

CHAPTER-1

Dupli Checker

PLAGIARISM SCAN REPORT

0%
Plagiarised

100%
Unique

Date	2020-06-05
Words	427
Characters	2714

Content Checked For Plagiarism

A user review is a review conducted by a consumer and published to a review site following buying a product or the evaluation of a service. As, more and more individuals and organizations have become accustomed to consulting user generated reviews before making purchases or online bookings. Considering great commercial benefits, merchants, however, have tried to hire people to write undeserving positive reviews to advertise their products or services, and meanwhile post malicious negative reviews to defame those of their competitors. Those spam reviews and opinions, which are deliberately produced in order to promote or demote targeted product or services, are known as deceptive opinion spam. 2. TYPES OF SPAM REVIEWS There are 3 types of Opinion spam reviews-(1) Fake reviews, (2) Brand reviews,(3) Non –reviews. (1) Fake reviews-In this type of reviews the user doesn't have the experience of the product/services that they are writing about. There is an insidious agenda behind it, either to promote a product or defame it. (2) Brand reviews- This user reviews are solely written by user on the basis of prior experience with the product/services of the particular brand and has nothing to do with that particular product. Past experience is the key here. (3) Non reviews-This are the user reviews that doesn't contains any relevant information or sentiment with respect to either the brand or the product. They mostly consist of advertisement of their product or services. The first type of reviews are the most difficult to detect. Fake reviews are the worst type of advertisement as they directly impact the reputation of a product/service. Whereas the (2) and (3) type of reviews are quite rare and have hardly any effect on the sentiment of the customers. ? 3. MOTIVATION Hotels or restaurants are vulnerable to fake reviews, particularly if they are negative. While rare, it can happen, whether it's an unethical competitor or an individual who has decided to cause problems for a business. A study conducted by Bright Local research revealed 82% of consumers read a fake review. Among youngsters the distribution was even higher. Allowing customers to be exposed to an increasing number of fake reviews, it's perfectly highly likely that we'll soon begin to see trustfulness in peer-to-peer recommendations being eroded. Therefore they should have a detection algorithm which filters out the spam reviews from the genuine reviews and improve the customer services. 4. OBJECTIVE OF THE PROJECT 1. Extraction of text features from the review text. 2. Application of suitable machine learning algorithm on the extracted features and increase the accuracy of the classification models.

Matched Source

No plagiarism found

Check By: Dupli Checker

CHAPTER-2

Dupli Checker

PLAGIARISM SCAN REPORT

0%
Plagiarised

100%
Unique

Date	2020-06-05
Words	383
Characters	2389

Content Checked For Plagiarism

1. LITERATURE REVIEW The first method for review spam detection was proposed in 2007 by Jindal and Liu .This paper was followed by [1] and [2] in which initial ideas were further investigated. Jindal et al. demonstrated that on data collection of 5.18 million reviews and 2.14 million reviewers from amazon (a e-commerce platform) showed duplication of reviews by using the method of spam detection. [2]. According to Dixit et al. he proposed that spam review can be divided into three cluster, (1) Reviews on Brands – failure of reviewing of the product as the reviews were only directed towards the seller or the brand, (2) Deceptive Reviews – was one of the main focus of this paper, and (3) Non-Reviews – comments that were majorly either irrelevant to the product or advertisements. [Dixit S, Agrawal AJ. Survey on review spam detection. Int J Comput Commun Technol ISSN (PRINT). 2013 Jun;4:0975-7449.]: Fei et al. [3]observed that using only word association features like n-grams alone proved inadequate for machine learning algorithms when learners were trained using synthetic fake reviews, since the features being created were not present in real-world fake reviews. Ott et al. [4] Tausczik et al. employed psycholinguistic features on the basis of features generated by LIWC [5]combined with standard word and Part of Speech (POS) n- gram features. Mukherjee et al. [6]extend that work including also style and POS based features, such as deep syntax and POS sequence patterns. The techniques that were supervised like Liu et al.[7] used a Bayesian approach and laid out a clustering problem with opinion spam sensing. Li et al. [4] 2. SYNOPSIS OF LITERATURE REVIEW From the Literature survey, it was understood on which dataset the project would be worked on. Reviews have been broadly classified in 3 categories- 1.Untruthful Reviews, 2.Non reviews, and 3.Reviews on brand. All of papers focused on Review centric features like Bag of Words or Parts of speech and found that the accuracy of the model with these features were quite poor. Therefore they tried to implement other stylometric features like length of words and sentences. Additional features such as such as Maximum content similarity can be employed provided the dataset contains information regarding the reviewer. Naive Bayes being the choice of algorithm they wanted to implement in supervised dataset.

Matched Source

No plagiarism found

Check By: Dupli Checker

CHAPTER-3

PLAGIARISM SCAN REPORT

2%
Plagiarised

98%
Unique

Date	2020-06-05
Words	974
Characters	6102

Content Checked For Plagiarism

1. DESCRIPTION OF THE FLOWCHART The Fig 3.1 represent the process in which the the classification model training and testing is done.At first the review text are extracted and then preprocessed to remove all the unwanted or non essential words,punctuations and other operations.Next step is feature extraction where the textual features are converted to vectorial representation.Then the generated matrix is then splitted to 2 parts i.e training and testing set.Model building is done on the training set with a choosen Machine learning algorithm.The build model is then tested on the testing set whose classification performance can be examined using a confusion matrix. 2. CHOICE OF PROGRAMMING LANGUAGE Till mid-term all the programming was done on R primarily because it has great visualization tools called ggplot2 and most of the report was based on data visualization .But python was choosen later in the project because of NLTK Toolkit it provides.It is easier to use for someone who is beginner at Natural Language processing to implement machine learning algorithm on text data. Numpy also provides a great help in representing the data into vectors and matrices.? 3. SELECTING THE RIGHT DATASET Data used in training and testing systems for spam detection comes from myriad of sources. Lack of standard datasets for this type of problem makes training of model under supervised learning algorithm difficult. The dataset used for this project consist of truthful and deceptive opinion reviews from 20 Chicago hotels collected from AMT and Yelp. It has been obtained from https://www.kaggle.com/rtatman/deceptive-opinion-spam-corpus 3.1 Reading the dataset. This dataset contains: 1. 400 truthful positive reviews from Trip Advisor 2. 400 deceptive positive reviews from Mechanical Turk. 3. 400 truthful negative reviews from Expedia, Hotels.com, Priceline, Trip Advisor and Yelp 4. 400 deceptive negative reviews from Mechanical Turk. Attributes of the dataset are- 1. Label- Describes if the Review text is classified as "truthful" or "deceptive". 2. Hotel- Describes the Name of the Hotel for which the Review was written. 3. Polarity- Describes the tone of the Review Text- "Positive" or "Negative" 4. Source- Mentions the website from which the Review text was extracted. 5. Review text- A textual representation of reviewer sentiment, emotions. ? The dimension of the dataset is -1600 rows and 5 columns. Following are the pre-processing steps that are done- 1. Removal of Stop words- Stop words are frequent words in any language which may not be useful or bring any valuable information in text .eg-"a","an","the". 2. Stemming- Stemming is the process of reducing words to their word stem; base or root, form generally a written word form e.g.- Consulting &Consultant have the same root word Consult. 3. Tokenization- Tokenization is the process of splitting a string, text into a list of words. 4. Parts of speech tagging (POS)- POS tags are labels given to words on basis of their context. This process includes tagging a word based on its definition and relationships with adjacent words. Words then are marked as Noun, Pronouns, etc. This information is collected and with additional features fed into a machine learning algorithm. 5. ?Removal of Numbers and Punctuation 4. FEATURE EXTRACTION Features that can be used to spam detection based on review content are as follows. Review centric features- 4.1 Bag of Words (BOW):Is a features extraction technique that is used for training machine learning algorithms. It creates vocabulary of all the distinct words occurring in all the reviews in the training set. Each unique word in the corpus is represented as a feature vector based on their occurrences in each of the review. Here each row index represent a particular review and each column consist of a unique word represent as number. 1 represent that a particular word occurred once in that particular review. 0 means absence. 4.2 Term frequency Inverse Document frequency (TFIDF) is a technique used to help compensate for words found relatively often in different reviews which makes it hard to distinguish between the reviews because they are too common. The Greater the frequency of a word in a review, the more important it is to the review. However the measurement is offset by the review size- the total number of words, the review contains- and by how often the word appears in other reviews. Similar to BOW each row index represents a review and each column represents a particular word. ? 4.3 Parts of

Dupli Checker

PLAGIARISM SCAN REPORT

4%
Plagiarised

96%
Unique

Date	2020-06-05
Words	510
Characters	3145

Content Checked For Plagiarism

4.6 Edit Distance(ED) is a way of quantifying how dissimilar two strings (e.g. words) are to one another by adding the minimum number of operations required to convert one string into the other. Basically it is minimum number of changes a given string has to be done to convert into a target string. These changes can be implemented by 3 operations- Insert, Remove and Replace. Each of the operations is of same cost. ? Fig 3.10 Minimum Edit Distance Here each row and column index represents a review and their corresponding value represents the edit distance between them. The higher the edit distance the higher is the dissimilarity between two reviews. So minimum edit distance has been taken for each reviews and used as feature vector of our model. Fig 3.11 Total minimum edit distances in ham and spam data 5. SPLITTING THE DATA Set Division Guidelines for Prediction Study Design • For large sample sizes – 60% training – 20% test – 20% validation • For medium sample sizes (Our dataset falls in these range) – 60% training – 40% test 6. ALGORITHM SELECTION 6.1 Naïve Bayes Goal in selecting models is to avoid over fitting on training data and minimize error on test data. From, Literature survey, there are two algorithms which provided the highest accuracy for spam detection on this dataset. Naive-Bayes and later Logistic Regression .For Naïve Bayes the Mathematical equation is represented by- ? for predictors X1. . .Xm, we want to model P(Y = k | X1, . . ,Xm) however, if we make the assumption that all predictor variables are independent to each other, the quantity can be simplified to- Naïve Bayes function P(Y = k), is determined from the data to be some known quantity ?k (also known as prior probability). 6.2 Logistic Regression Algorithm is similar to linear regression, with the only difference being the y data which should contain integer values and represents class. Here the two classes are separate by a decision boundary (a linear line).It consist of sigmoid function and an error function. Less the value of cost function more accurate will be the model. Regularization parameter can be used to reduce over fitting on the training data. Fig 3.12 Sigmoid Function Fig 3.13 Cost Function of Logistic Regression ? 7. DIMENSION REDUCTION Principal Component analysis (PCA)-Constructing a classification model may not require every feature. Ideally we want to capture the most variation in data with the least number of variables.PCA is suited to do this and will help reduce number of features as well as reduce noise Linear discriminate analysis (LDA)-is a dimension reduction technique where a new feature space is found out in order to maximize the class separability in the data. Specifically, the model wants to find a linear combination of input features that achieves the maximum separation for data between classes and the minimum separation of samples within each class. 8. TRAINING THE MODEL Here we use scikitlearn package in Python which consist of many data pre-processing steps and machine learning model. Link for scikitlearn information page-https://scikit-learn.org/stable/ ?

Matched Source

Similarity 7%

Title: Practical Machine Learning Course Notes | Receiver Operating...

build the model/predictor on the remaining training data in each subset and applied to the test subset rebuild the data k times with the training and testIdeally we want to capture the most variation with the least amount of variables weighted combination of predictors may improve fit combination needs...

CHAPTER-4

Dupli Checker

PLAGIARISM SCAN REPORT

0%
Plagiarised

100%
Unique

Date	2020-06-05
Words	938
Characters	6119

Content Checked For Plagiarism

2.1 Accuracy of different models 2.1.1 Naïve Bayes 1. BOW–Matrix that contains mostly zero values are called sparse matrix. Features like bag of words generate a huge sparse matrix. So computation becomes difficult and time expensive. Also it reduces the accuracy of the model as observed from the graph from our data. So to increase the accuracy we reduce the matrix with sparsity of 0.9732 (min df =5).Therefore obtain an accuracy of 0.692. Table 4.1 Performance of Bow model Fig 4.1 Accuracy vs. Sparsity 2. TFIDF+BOW-By combining both the features the Bag of words and TFIDF ,now each of the words are represented by their weights in a particular review in addition to the occurrences. Accuracy increased to 0.693. Table 4.2 Performance of Bow+TFIDF model 3. POS+TFIDF+BOW-Parts of speech counts have been added to the feature vector which initially consist of TFIDF+BOW features. Accuracy improved to 0.695. Table 4.3 Performance of Bow+TFIDF+POS model 4. POS+TFIDF+BOW+RL-With the review length being added the dimensionality of the feature vector increased. Till now the feature vector primarily consist vector input from-1 to 1, but review length has very different set of input from-1 to 1. Since we are using Naïve Bayes algorithm, we can skip feature scaling as it is a graphical model based classifier. Accuracy improved significantly to 0.747. Table 4.4 Performance of Bow+TFIDF+POS+RL model 5. POS+TFIDF+BOW+RL+SP-The sentiment function in NLTK Toolkit will return a named tuple of form Sentiment (polarity, subjectivity). The polarity score is a float number within the range (-1.0, 1.0). we consider only the polarity score. Accuracy remains the same. So this feature might not be helpful and acts as noise. Table 4.5 Performance of Bow+TFID+POS+RL+SP model 6. POS+TFIDF+BOW+RL+SP+ED- Edit distance is new feature that is added on to the the feature vectors. Accuracy of the model is increased to 0.753. Table 4.6 Performance of Bow+TFIDF+POS+RL+SP+ED model 7. Model selection and Feature Combination Fig 4.2 Model Selection (Naïve Bayes) Here we observe that accuracy of our model continues to increase as the number of features in our model increases. But sensitivity of our model first increases and then reduces. Specificity first decreases then increases. Therefore the best model is the one which have a higher accuracy with reduced cases of false positives and false negatives i.e point of intersection of the three curves. In this case we observe from the plot that when number of features we used is four, the model performance is overall the best. Therefore for Naïve Bayes classifier a feature combination of (TFIDF+BOW+POS+RL) generates a good model. ? 2.1.2 Logistic regression For the sake of visualizing the decision boundary, we convert our large feature vector to 2 dimensions feature vector using PCA. Then the training set was used to build the model using Logistic Regression algorithm and then test on the testing set. Following is the plot of the decision boundary of our dataset. Here PC1 represents the first feature vector and PC2 represents the second feature vector. The separation boundary between red and green region is the linear decision boundary. 0 and 1 are the two classes i.e. Truthful and Deceptive. Fig 4.3 2D plot of decision boundary by Logistic Regression 1. BOW- Here the regularization parameter is set to one with number of iterations to be 100.Sparsity is 0.97. Accuracy is about 0.856. Table 4.7 Performance of Bow model 2. TFIDF+BOW-Same as before with regularization parameter set to one and number of iterations set to 100.Accuracy increases to 0.876. Table 4.8 Performance of TFIDF+Bow model ? 3. TFIDF+BOW+POS-Here the accuracy of model decreases in contrast to Naïve Bayes model. Therefore POS acts as a noise and is not helpful for classification. Accuracy is 0.835 Table 4.9 Performance of TFIDF+BOW+POS model 4. TFIDF+BOW+POS+RL-Accuracy increases to 0.842 Table 4.10 Performance of TFIDF+BOW+POS+RL model 5. TFIDF+BOW+POS+RL+SP-Same as Naïve Bayes, SP acts as a noise for both the classifier and therefore a bad feature vector altogether. Accuracy is 0.842. Table 4.11 Performance of Bow +TFIDF+POS+RL+SP model 6. TFIDF+BOW+POS+RL+SP+ED-Number of iterations have been increased to 1000 to reach the convergence of the cost function. Accuracy has increased to 0.846 Table 4.12 Performance of Bow+TFIDF+POS+RL+SP+ED model ? 7. Model selection and feature

CONCLUSSION

Dupli Checker

PLAGIARISM SCAN REPORT

0%
Plagiarised

100%
Unique

Date	2020-06-06
Words	170
Characters	1127

Content Checked For Plagiarism

1. Logistic Regression learns from the training data taking into account all the possible correlations between features. While in Naïve Bayes there is a perquisite assumption that all the features act independently to each other. Therefore Logistic Regression is more prone to overfitting the data than Naïve Bayes. This is the reason why there is higher accuracy in case of the former than the latter. 2. TFIDF+BOW+POS+RL feature combination generates best model with Naïve Bayes algorithm .TFIDF+BOW feature combination generates best model with Logistic Regression. 3. Though accuracy should increase with increase in features, we have seen that adding sentiment polarity on both Naïve Bayes and Logistic regression algorithms reduced the model performance. It is a noise and doesn't provide any meaningful information to our classifier irrespective of the classifier. 4. Dimension reduction such as LDA can be applied to generate a 1D feature vector from a large matrix. This generated feature can later be used for training models with Logistic Regression and Naïve Bayes which increases the accuracy significantly.

Matched Source

No plagiarism found

Check By: Dupli Checker