REVIEW CLASSIFICATION IN E-COMMERCE USING SENTIMENT ANALYSIS

SUBHALAXMI ANNAMREDDI

ABSTRACT: In this project we consider and analyze the reviews to e-commerce websites given by the customers. Here when a customer reviews about a product we analyze the review whether the word is good or bad using data mining based sentiment analysis. Basing on the score obtained the ratings are assigned. The reviews can be given not only to products but also to registered shops i.e. when a customer bought something he can review his experience with the store. All this process is helpful to analyze the income in e-commerce sites and predict the future income of that site or product. This project can be used as a backend service and could be customized like advertising the shops or hotels with offers based on customer interests. To classify the text random forest classifier is used.

Keywords- prediction, data mining, sentiment analysis, random forest.

1.Introduction:

1.1 Project Overview

In this project we consider and analyze the reviews to e-commerce websites given by the customers. Here when a customer reviews about a product we analyze the review whether the word is good or bad using data mining based sentiment analysis. Basing on the score obtained the ratings are assigned. The reviews can be given not only to products but also to registered shops i.e. when a customer bought something he can review his experience with the store. All this process is helpful to analyze the income in e-commerce sites and predict the future income of that site or product. This project can be used as a backend service and could be customized like advertising the shops or hotels with offers based on customer interests. To classify the text random forest classifier is used.

The review data is encoded into vector format (i.e., 0’s and 1’s). The tf-idf values are generated using those encoded values and random forest algorithm is applied to generate the opinion of the review. The products are classified and plotted onto a bar graph.

1.2. Project Deliverables

A Project Deliverable is a product or service that a project produces for its customer, client, or project sponsor. It is the product or service that the project “delivers” to its stakeholders.It can be tangible or intangible, for example, a contractor who is hired to provide a training course provides the course itself back to their client as the project’s deliverable.

In our project we analyze the reviews of customer to create a classifier to predict the polarity of user reviews. We will create a confusion matrix to compare between actual observations and predicted observations.

1.3. Project scope

People have always had an interest in what people think, or what their opinion. Since the inception of the internet, increasing numbers of people are using websites and services to express their opinion.

With social media channels such as Face book, LinkedIn, and Twitter, it is becoming feasible to automate and gauge what public opinion is on a given topic, news story, product, or brand. Opinions that are mined from such services can be valuable. Datasets that are gathered can be analyzed and presented in such a way that it becomes easy to identify if the online mood is positive, negative .

This allows individuals or business to be proactive as opposed to reactive when a negative conversational thread is emerging. Alternatively, positive sentiment can be identified thereby allowing the identification of product advocates or to see which parts of a business strategy are working.

Sentiment analysis has many applications and benefits to your business and organization. It can be used to give your business valuable insights into how people feel about your product brand or service.

Currently, sentiment analysis is a topic of great interest and development since it has many practical applications. Since publicly and privately available information over Internet is constantly growing, a large number of texts expressing opinions are available in review sites, forums, blogs, and social media.

Sentiment analysis can be applied at different levels of scope:

Document level sentiment analysis obtains the sentiment of a complete ocument or paragraph.

Sentence level sentiment analysis obtains the sentiment of a single sentence.

Sub-sentence level sentiment analysis obtains the sentiment of sub-expressions within a sentence.

It's estimated that 80% of the world's data is unstructured and not organized in a pre-defined manner. Most of this comes from text data, like emails, support tickets, chats, social media, surveys, articles, and documents. These texts are usually difficult, time- consuming and expensive to analyze, understand, and sort through.

Sentiment analysis systems allows companies to make sense of this sea of unstructured text by automating business processes, getting actionable insights, and saving hours of manual data processing, in other words, by making teams more efficient.

2.1. **Existing system**

The customer review is important to improve service for company, which have both close opinion and open opinion. The open opinion means the comment as text which shows emotion and comment directly from customer. However, the company has many contents or group to evaluation themselves by rating and total rating for a type of services which there are many customer who needs to review. The problem is some customers given rating contrast with their comments. The other reviewers must read many omments and comprehensive the comments that are different from the rating. Therefore, this paper proposes the analysis and prediction rating from customer reviews who commented as open opinion using probability&#39;s classifier model. The classifier models are used case study of customer review&#39;s hotel in open comments for training data to classify comments as positive or negative called opinion mining. In addition, this classifier model has calculated probability that shows value of trend to give the rating using naive bayes techniques with decision tree Techniques

NaIve bayes is an algorithm of probability based on Bayes theorem of learning. It aims to create a model in the form of probability.

The advantage of naIve bayes is an effective method which is easy processing. The probability of the classification data with prior knowledge is denoted by P(ail Vj), where ai refers to the attribute i and Vj refers to class label j. Therefore, the classification has been calculated for this probability. The highest probability of ai is depended on Vj for each class is trend to answer of classification. The range of probability is between 0 and 1 as

VNB = arg max P (v j) *TI =1 P (a i I v j)

- Decision Tree(C4.5) The decision tree learning was proposed as a model of data classification for a class label, which called ID3 and developed to C4.5. In addition, decision tree is clearly represented through a tree diagram. It starts from the fIrst node is a root node. The root node selects an attribute as words in opinion from the best value of measurement. Each attribute has its own values i.e. true/false, which are separated by branch links composed of original attributes. At the end, the data reveals a class which represents a leaf node (i.e. positive/negative). The advantage of the decision tree is for ordering attributes that are the best measurement

2.1.1. **Challenges:**

- The first disadvantage is that the Naive Bayes classifier makes a very strong assumption on the shape of your data distribution, i.e. any two features are independent given the output class.
- Another problem happens due to data scarcity. For any possible value of a feature, you need to estimate a likelihood value by a frequentist approach. This can result in probabilities going towards 0 or 1, which in turn leads to numerical instabilities and worse results.
- Naïve bayes classifier is relatively slower for larger data sets and so is the accuracy falls down for the larger data sets.

2.2. **Proposed system:**

In this project we consider and analyze the reviews to e-commerce websites given by the customers. Here when a customer reviews about a product we analyze the review whether the word is good or bad using data mining based sentiment analysis. Basing on the score obtained the ratings are assigned. The reviews can be given not only to products but also to registered shops i.e. when a customer bought something he can review his experience with the store. All this process is helpful to analyze the income in e-commerce sites and predict the future income of that site or product.

This project can be used as a backend service and could be customized like advertising the shops or hotels with offers based on customer interests. To classify the text random forest classifier is used.

Random Forests are an improvement over bagged decision trees.

Random forest changes the algorithm for the way that the sub-trees are learned so that the resulting predictions from all of the subtrees have less correlation.

The number of features that can be searched at each split point (m) must be specified as a parameter to the algorithm. You can try different values and tune it using cross validation.

For classification a good default is: m = sqrt(p)

For regression a good default is: m = p/3

Where m is the number of randomly selected features that can be searched at a split point and p is the number of input variables.

The performance of each model on its left out samples when averaged can provide an estimated accuracy of the bagged models. This estimated performance is often called the OOB estimate of performance.

2.2.1. Advantages:

One big advantage of random forest is, that it can be used for both classification and regression problems, which form the majority of current machine learning systems.

Random Forest is considered as a very handy and easy to use algorithm, because it's default hyper parameters often produce a good prediction result. The number of hyper parameters is also not that high and they are straightforward to understand.

□ It gives better results with the increasing number of examples.

□ It is robust against over fitting at least with my experiences and the claims of the creator Leo Breiman and Adele Cutler.

□ It might be used for clustering, statistical inference and feature selection as well Works good with numerical, categorical data.

And on top of that, they can handle a lot of different feature types, like binary, categorical and numerical.

3.1. System architecture

An architectural design is the design of the entire software system; it gives a high-level overview of the software system, such that the reader can more easily follow the more detailed descriptions in the later sections. It provides information on the decomposition of the system into modules (classes), dependencies between modules, hierarchy and partitioning of the software modules.

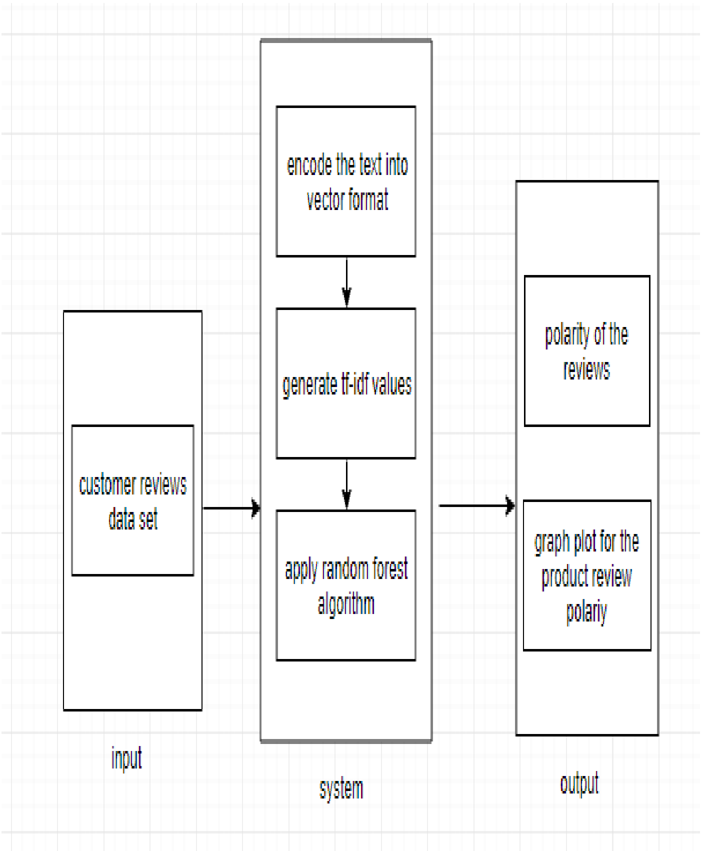

customer reviews data set
input
encode the text into vector format
generate tf-idf values
apply random forest algorithm
system
polarity of the reviews
graph plot for the product review polariy
output

Figure: **System Architecture**

3.2.1. Algorithm description

Random forest classifier:

Random Forest is a flexible, easy to use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time. It is also one of the most used algorithms, because it's simplicity and the fact that it can be used for both classification and regression tasks.

Random Forest is a flexible, easy to use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time. It is also one of the most used algorithms, because it's simplicity and the fact that it can be used for both classification and regression tasks.

To say it in simple words: Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction.

Random Forest adds additional randomness to the model, while growing the trees. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. This results in a wide diversity that generally results in a better model.

Therefore, in Random Forest, only a random subset of the features is taken into consideration by the algorithm for splitting a node. You can even make trees more random, by additionally using random thresholds for each feature rather than searching for the best possible thresholds

Sklearn provides a great tool for this, that measures a features importance by looking at how much the tree nodes, which use that feature, reduce impurity across all trees in the forest. It computes this score automatically for each feature after training and scales the results, so that the sum of all importance is equal to 1.

The Hyperparameters in random forest are either used to increase the predictive power of the model or to make the model faster.

Increasing the Predictive Power

- There is the **„n_estimators“** hyperparameter, which is just the number of trees the algorithm builds before taking the maximum voting or taking averages of predictions.
- Another important hyperparameter is **„max_features“**, which is the maximum number of features Random Forest considers to split a node.
- The last important hyper-parameter is **„min_sample_leaf “**. This determines the minimum number of leafs that are required to split an internal node.

Random Forest pseudocode:

1. Randomly select **"k"** features from total **"m"** features.
    1. Where **k << m**
2. Among the **"k"** features, calculate the node **"d"** using the best split point.
3. Split the node into **daughter nodes** using the **best split**.
4. Repeat **1 to 3** steps until "l" number of nodes has been reached.
5. Build forest by repeating steps **1 to 4** for "n" number times to create **"n" number of trees**.

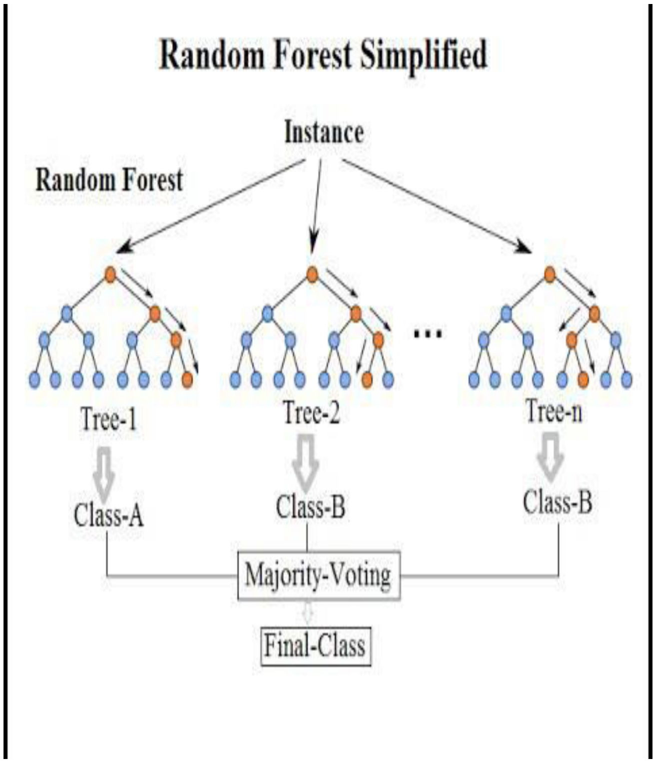


Figure 5.4 Random forest

TF-IDF:

Tf-idf stands for *term frequency-inverse document frequency*, and the tf-idf weight is a weight often used in information retrieval and text mining.

This weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus.

TF-IDF score represents the relative importance of a term in the document and the entire corpus. TF-IDF score is composed by two terms: the first computes the normalized Term Frequency (TF), the second term is the Inverse Document Frequency (IDF), computed as the logarithm of the number of the documents in the corpus divided by the number of documents where the specific term appears.

TF(t) = (Number of times term t appears in a document) / (Total number of terms in the document)

IDF(t) = log_e(Total number of documents / Number of documents with term t in it)

TF-IDF Vectors can be generated at different levels of input tokens (words, characters, n-grams)

a. Word Level TF-IDF : Matrix representing tf-idf scores of every term in different documents

b. N-gram Level TF-IDF : N-grams are the combination of N terms together. This Matrix representing tf-idf scores of N-grams

c. Character Level TF-IDF : Matrix representing tf-idf scores of character level n-grams in the corpus

LDA: LDA or latent Dirichlet allocation is a “generative probabilistic model” of a collection of composites made up of parts. In terms of topic modeling, the composites are documents and the parts are words and/or phrases (n-grams).

The probabilistic topic model estimated by LDA consists of two tables.

- The first table describes the probability or chance of selecting a particular part when sampling a particular topic (category).

- The second table describes the chance of selecting a particular topic when sampling a particular document or composite.

Using LDA is a way of soft clustering your composites and parts.

If you choose the number of topics to be less than the documents, using LDA is a way of reducing the dimensionality (the number of rows and columns) of the original composite versus part data set.

Algorithm steps:

Step 1: Collect the reviews from customers.

Step 2: Encode the text reviews into vector format.

Step 3: Calculate the tf-idf values to each review data.

Step 4: Apply random forest algorithm and display the polarity of reviews.

Step 5: Display the word cloud format of mostly bought products

Step 6: Display graph plot for number of positive and negative reviews of individual product.

4.1 Technology Description

4.1.1. Anaconda Navigator:

Anaconda Navigator is a desktop graphical user interface (GUI) included in Anaconda distribution that allows you to launch applications and easily manage conda packages, environments and channels without using command-line commands. Navigator can search for packages on Anaconda Cloud or in a local Anaconda Repository. It is available for Windows, macOS and Linux.

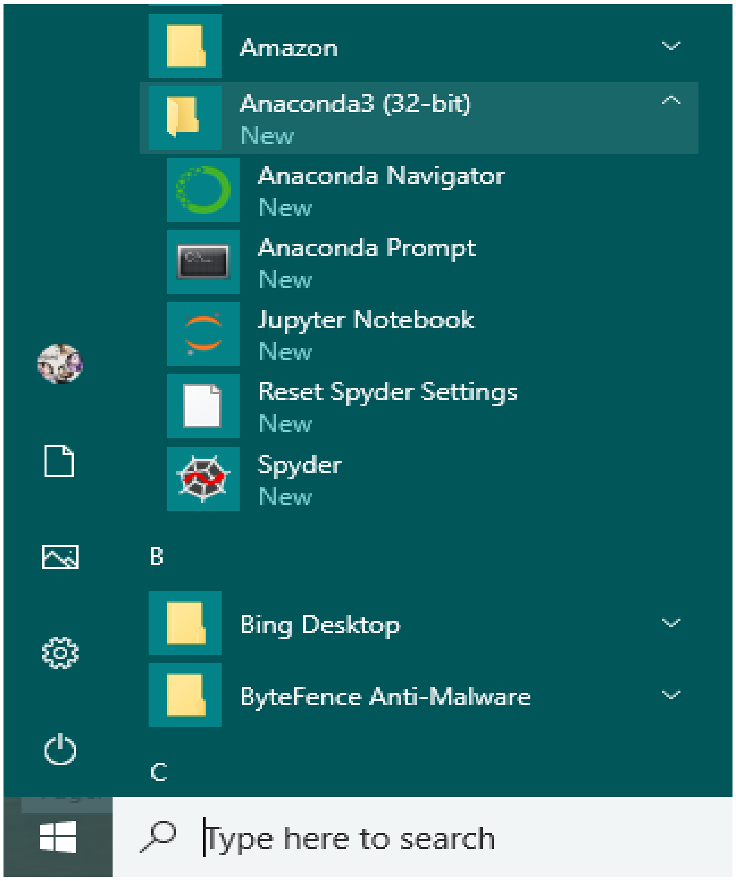


Figure 6.1 Anaconda Navigator

In order to run, many scientific packages depend on specific versions of other packages. Data scientists often use multiple versions of many packages, and use multiple environments to separate these different versions.

5.1. Testing

Testing is the major quality control measure employed for software development. Its basic function is to detect errors in the software. During requirement analysis and design, the output is a document which is usually textual and non-textual. After the coding phase, computer programs are available that can be executed for testing purpose. This implies that testing has to uncover errors introduced during coding phases. Thus, the goal of testing is to cover requirement, design, or coding errors in the program. The purpose is to exercise the different parts of the module code to detect coding errors. After this, the modules are gradually integrated into subsystems, which are then integrated themselves to eventually form the entire system. During the module integration, testing is performed. The goal is to detect designing errors, while focusing the interconnection between modules. After the system was put together, system testing is performed. Here the system is tested against the system requirements to see if all requirements were met and the system performs as specified by the requirements. Finally, testing is performed to demonstrate to the client for the operation of the system.

For the testing to be successful, proper selection of the test case is essential. There are two different approaches for selecting test case. The software or the module to be tested is treated as a black box, and the test cases are decided based on the specifications of the system or module. For this reason, this form of testing is also called “black box testing”.

The focus here is on testing the external behavior of the system. In structural testing, the test cases are decided based on the logic of the module to be tested. A common approach here is to achieve some type of coverage of the statements in the code. The two forms of testing are complementary: one tests the external behavior, the other tests the internal structure. Often structural testing is used for lower levels of testing, while functional testing is used for higher levels.

Testing is an extremely critical and time-consuming activity. It requires proper planning of the overall testing process. Frequently the testing process starts with the test plan. This plan identifies all testing related activities that must be performed and specifies the schedule, allocates the resources, and specifies guidelines for testing. The test plan specifies conditions that should be tested; different units to be tested, and the manner in which the module will be integrated together. Then for different test unit, a test case specification document is produced, which lists all the different test cases, together with the expected outputs, that will be used for testing. During the testing of the unit the specified test cases are executed and the actual results are compared with the expected outputs. The final output of the testing phase is the testing report and the error report, or a set of such reports. Each test report contains a set of test cases and the result of executing the code with the test cases. The error report describes the errors encountered and the action taken to remove the error.

Testing approach

Testing is a process, which reveals the errors in a program. It is the major quality measure employed during software development. During testing, the program is executed with a set of conditions known as test cases and the output is evaluated to determine whether the program is performing as expected. In order to make sure that the system does not have errors, the different levels of testing strategies are applied at differing phases of software development are as follows.

5.1.1 Unit Testing

Unit Testing is done on individual modules as they are completed and become executable. It is confined only to the designer's requirements.

In our project (REVIEW CLASSIFICATION IN E-COMMERCE USING SENTIMENT ANALYSIS). we are using RANDOM FOREST algorithm to classify and predict the reviews of the data.

5.1.2 Each module can be tested using the following two strategies

5.1.2.1 Black Box Testing

Internal system design is not considered in this type of testing. Tests are based on the requirements and the functionality. This testing is used to find the errors in the following categories:

- Incorrect or missing functions
- Interface errors
- Errors in data structure
- Performance errors
- Initialization and termination errors.

In this testing, only the output is checked for correctness but the logical flow of the data is not checked.

5.1.2.2 White Box Testing

This testing is based on the knowledge of the internal logic of an application’s code. Also known as Glass box Testing. Internal software and code working should be known for this type of testing. Tests are based on coverage of code, statements, etc. It is used to generate the test cases in the following cases:

- Guarantee that all the independent paths have been executed.
- Execute all the logical decisions on their true and false sides.
- Execute all the loops at their boundaries and within their operational
- Execute the internal data structures to ensure their validity.

5.1.3 Integration Testing

Integration testing ensures that the software and the subsystems work together as a whole. It tests the interface of all the modules to make sure that the modules behave properly or not when integrated together.

5.1.4 System Testing

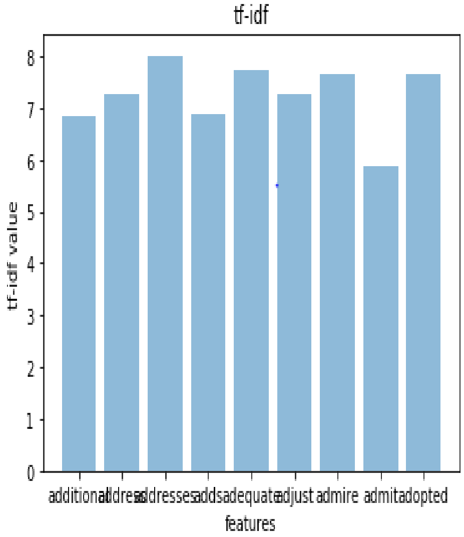
It involves in-house testing of the entire system before the delivery to the user. Its aim is to satisfy the user and the system that meets all the requirements of the client's specifications.

Test cases

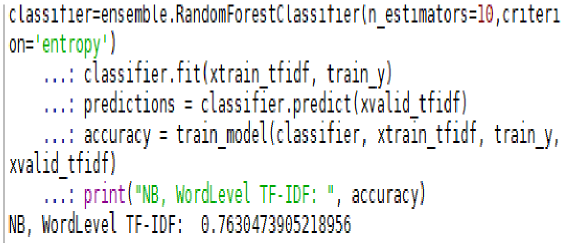
Table7 .1 Test Cases

S No.	Description	Expected Value	Input	Actual Value	Result
1.	Given review is positive	Positive	Positive review	Positive	Pass
2.	Given review is negative	Negative	Negative review	Negative	Pass
3.	Given review is positive (not bad)	Positive	Positive review	Positive	Pass
4.	Given review is positive	Positive	Positive review	Negative	Fail
5.	Given review is negative (not good)	Negative	Negative review	Positive	Fail

6.1. Tf-idf values:



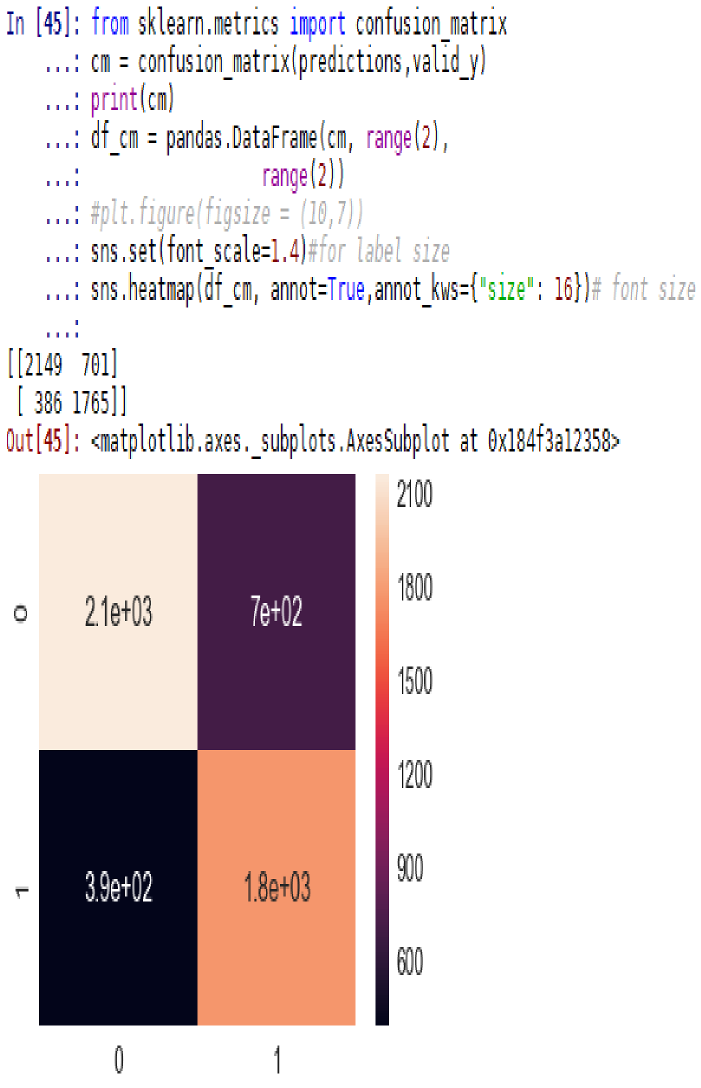
6.2. Classification:



```
classifier=ensemble.RandomForestClassifier(n_estimators=10,criteri
on='entropy')
    ...: classifier.fit(xtrain_tfidf, train_y)
    ...: predictions = classifier.predict(xvalid_tfidf)
    ...: accuracy = train_model(classifier, xtrain_tfidf, train_y,
xvalid_tfidf)
    ...: print("NB, WordLevel TF-IDF: ", accuracy)
NB, WordLevel TF-IDF:  0.7630473905218956
```

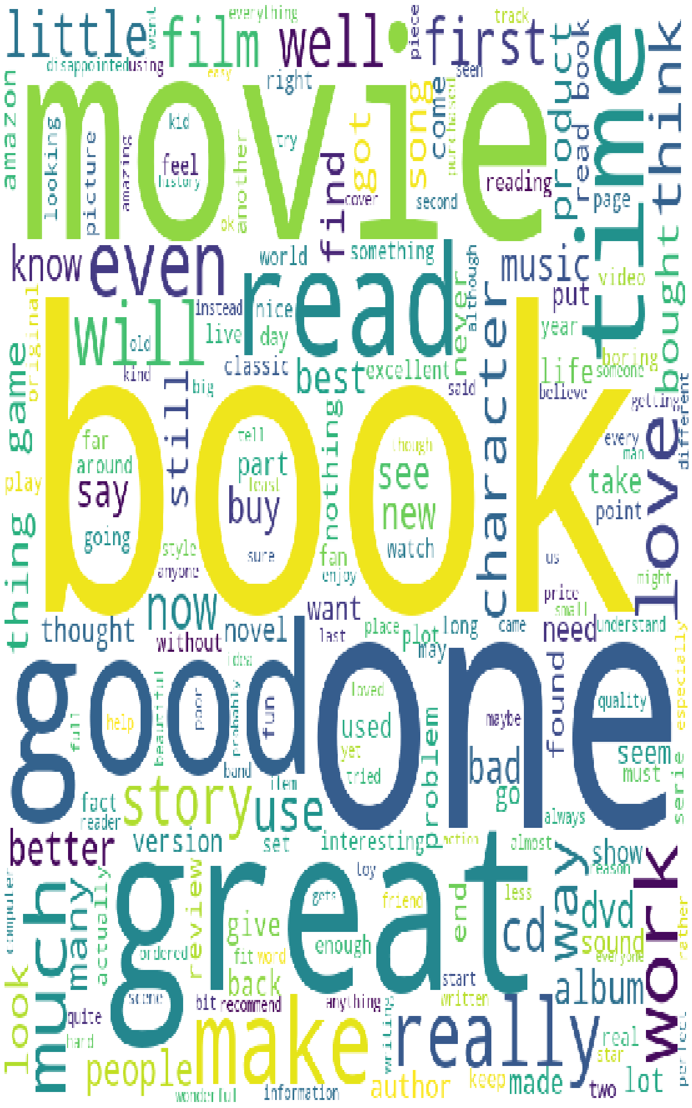
6.3. Confusion matrix:

```
In [22]: from sklearn.metrics import
confusion_matrix
    ...: cm =
confusion_matrix(predictions,valid_y)
    ...: print(cm)
    ...:
[[2149  701]
 [ 386 1765]]
```

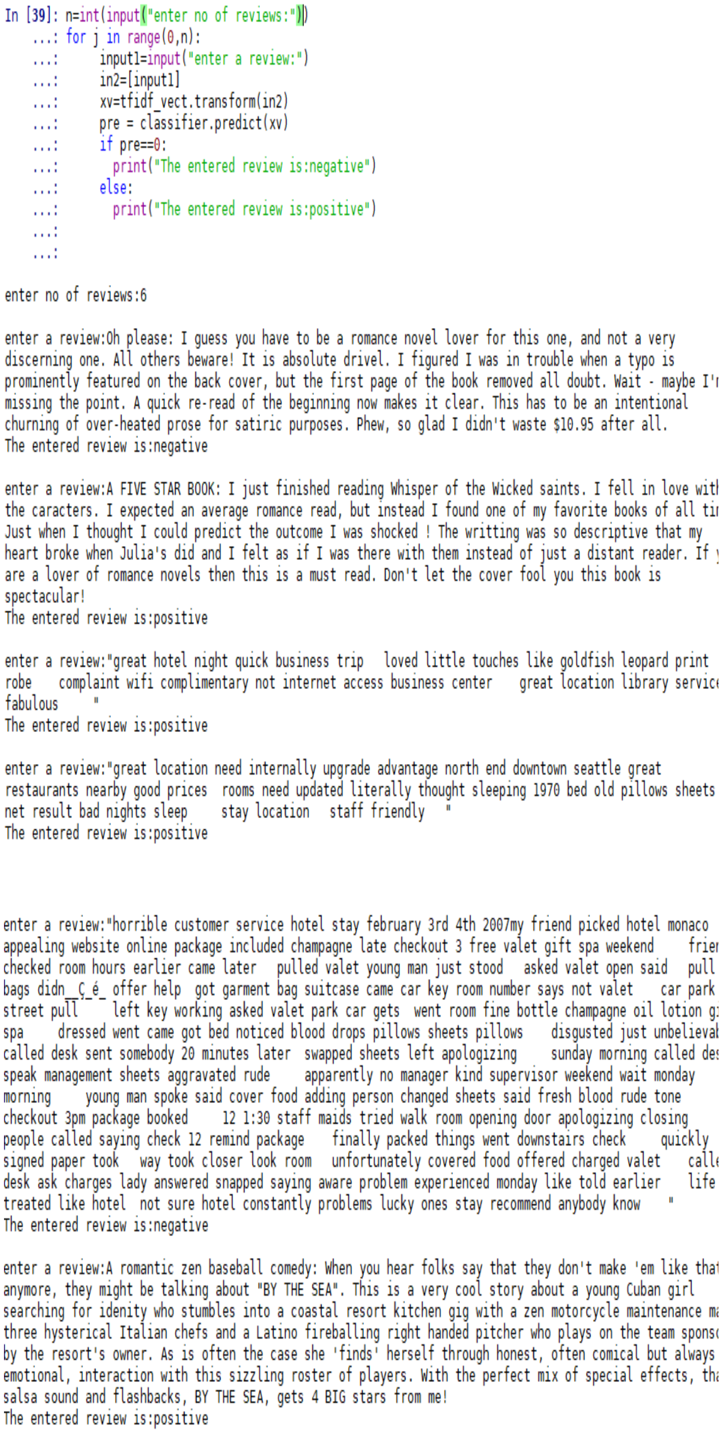


```
In [45]: from sklearn.metrics import confusion_matrix
    ...: cm = confusion_matrix(predictions,valid_y)
    ...: print(cm)
    ...: df_cm = pandas.DataFrame(cm, range(2),
    ...:                   range(2))
    ...: #plt.figure(figsize = (10,7))
    ...: sns.set(font_scale=1.4)#for label size
    ...: sns.heatmap(df_cm, annot=True,annot_kws={"size": 16})# font size
    ...:
[[2149  701]
 [ 386 1765]]
Out[45]: <matplotlib.axes._subplots.AxesSubplot at 0x184f3a12358>
```

6.4. Word cloud representation:

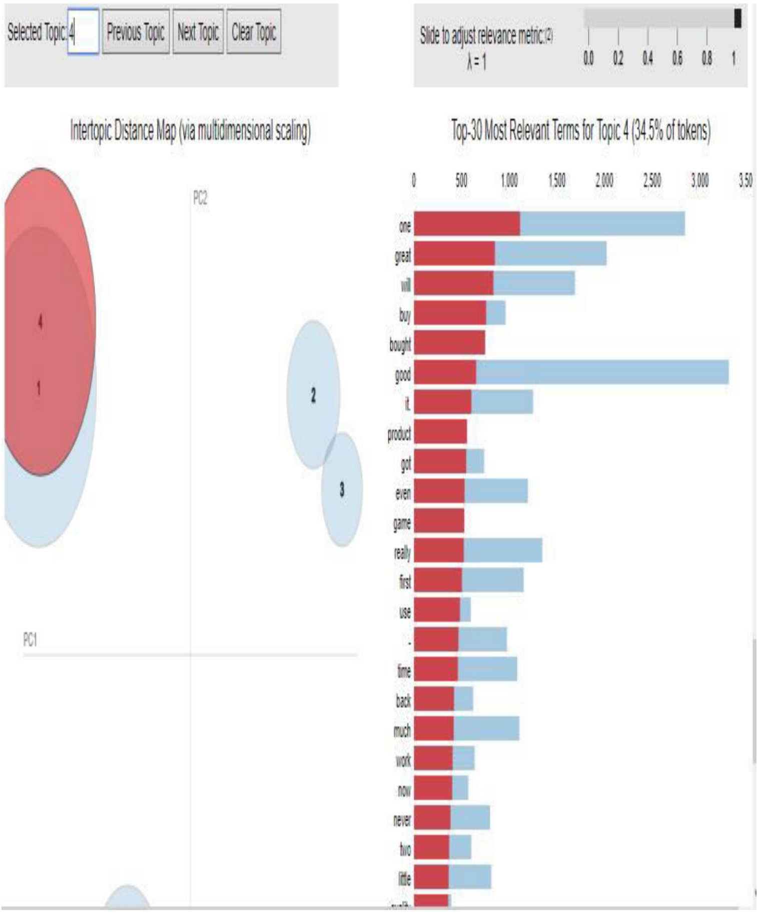


6.5. Opinion of a review:



```
In [39]: n=int(input("enter no of reviews:"))
    ...: for j in range(0,n):
    ...:     input1=input("enter a review:")
    ...:     in2=[input1]
    ...:     xv=tfidf_vect.transform(in2)
    ...:     pre = classifier.predict(xv)
    ...:     if pre==0:
    ...:       print("The entered review is:negative")
    ...:     else:
    ...:       print("The entered review is:positive")
    ...:
    ...:

enter no of reviews:6

enter a review:Oh please: I guess you have to be a romance novel lover for this one, and not a very
discerning one. All others beware! It is absolute drivel. I figured I was in trouble when a typo is
prominently featured on the back cover, but the first page of the book removed all doubt. Wait - maybe I'
missing the point. A quick re-read of the beginning now makes it clear. This has to be an intentional
churning of over-heated prose for satiric purposes. Phew, so glad I didn't waste $10.95 after all.
The entered review is:negative

enter a review:A FIVE STAR BOOK: I just finished reading Whisper of the Wicked saints. I fell in love wit
the caracters. I expected an average romance read, but instead I found one of my favorite books of all ti
Just when I thought I could predict the outcome I was shocked ! The writting was so descriptive that my
heart broke when Julia's did and I felt as if I was there with them instead of just a distant reader. If
are a lover of romance novels then this is a must read. Don't let the cover fool you this book is
spectacular!
The entered review is:positive

enter a review:"great hotel night quick business trip   loved little touches like goldfish leopard print
robe    complaint wifi complimentary not internet access business center    great location library servic
fabulous    "
The entered review is:positive

enter a review:"great location need internally upgrade advantage north end downtown seattle great
restaurants nearby good prices  rooms need updated literally thought sleeping 1970 bed old pillows sheets
net result bad nights sleep     stay location   staff friendly   "
The entered review is:positive



enter a review:"horrible customer service hotel stay february 3rd 4th 2007my friend picked hotel monaco
appealing website online package included champagne late checkout 3 free valet gift spa weekend     frie
checked room hours earlier came later   pulled valet young man just stood   asked valet open said   pull
bags didn__Ç_é_ offer help  got garment bag suitcase came car key room number says not valet    car park
street pull     left key working asked valet park car gets  went room fine bottle champagne oil lotion g
spa     dressed went came got bed noticed blood drops pillows sheets pillows    disgusted just unbelieva
called desk sent somebody 20 minutes later  swapped sheets left apologizing     sunday morning called de
speak management sheets aggravated rude     apparently no manager kind supervisor weekend wait monday
morning     young man spoke said cover food adding person changed sheets said fresh blood rude tone
checkout 3pm package booked     12 1:30 staff maids tried walk room opening door apologizing closing
people called saying check 12 remind package    finally packed things went downstairs check     quickly
signed paper took   way took closer look room   unfortunately covered food offered charged valet    call
desk ask charges lady answered snapped saying aware problem experienced monday like told earlier    life
treated like hotel  not sure hotel constantly problems lucky ones stay recommend anybody know    "
The entered review is:negative

enter a review:A romantic zen baseball comedy: When you hear folks say that they don't make 'em like tha
anymore, they might be talking about "BY THE SEA". This is a very cool story about a young Cuban girl
searching for idenity who stumbles into a coastal resort kitchen gig with a zen motorcycle maintenance m
three hysterical Italian chefs and a Latino fireballing right handed pitcher who plays on the team spons
by the resort's owner. As is often the case she 'finds' herself through honest, often comical but always
emotional, interaction with this sizzling roster of players. With the perfect mix of special effects, th
salsa sound and flashbacks, BY THE SEA, gets 4 BIG stars from me!
The entered review is:positive
```

6.6. Plotting most relevant terms:



Today, the sentiment analysis has become an important task and also spread widely used in any country and any language. This makes many approaches have been proposed in different type of languages.

Natural language processing tells us the opinion of the customer about a product or about a seller through the customer written reviews.

This helps the customers to have better options in choosing a product.

These statistics are useful for the seller to know the opinions of the customers about his products. So that the seller or the owner can rectify the problems with their products and they can also predict their future sales.

[1] N. Kumari and S. N. Singh, Sentiment analysis on E-commerce application by using opinion mining, Proceeding of the 6th International Conference-Cloud System and Big Data Engineering(Confluence), 2016, pp. 320-325.

[2] Anitha, N., Anitha, B., & Pradeepa, S. (2013). Sentiment Classification Approaches–A Review. International Journal of Innovations in Engineering and Technology (IJIET), 3(1), 22-31.

[3] T. Chumwatana, Using sentiment analysis technique for analyzing Thai customer satisfaction from social media. Proceeding of the 5th International Conference on Computing and Informatics, 2015, pp.659- 664.

[4] S.Ahmed and A.Danti, A novel Approach for sentimental analysis and opinion mining based on sentiwordnet using web data. Proceeding of International Conference on Trends in Automation, Communications and Computing Technology, 2015, pp.1-5.

[5] R.K. Bakshi, N. Kaur, R. Kaur and G.Kaur, Opinion mining and sentiment analaysis, Proceeding of the 3rd International Conference on Computing for Sustainable Global Development, 2016, pp. 452-455.

[6] P.Barnaghim, 1.G. Breslin and P. Ghaffari, Opinion mining and sentiment polarity on Twiiter and correlation between events and sentiment, Proceeding of the 2nd International Conference on Big Data Computing Service and Application, 2016, pp. 52-57.

[7] Kumar Singh, P., Sachdeva, A., Mahajan, D., Pande, N., & Sharma, A. (2014, September). An approach towards feature specific opinion mining and sentimental analysis across e-commerce websites. In Confluence The Next Generation Information Technology Summit (Confluence), 2014 5th International Conference- (pp. 329-335). IEEE.

[8] S.Atia and K. Shaalan, Increasing the accuracy of opinion mining in Arabic. Proceeding of the 1st International conference on Arabic computing linguistics, 2015, pp.l 06-113.