

**ABSTRACT**

In the modern era of computer science where an individual’s decisions is influenced by online reviews of peers and strangers, it is important for businesses to find the general sentiment of the crowd towards their business and their opinions.

A Natural Language Processing task called Sentiment Analysis wherein the semantic orientation of a text review is judged to be either positive or negative. A major challenge in Sentiment Analysis is the identiﬁcation of aspects towards which the opinion is expressed. Aspect based Sentiment Analysis can be divided into two tasks. The ﬁrst part being the extraction of the aspect term from a sentence and secondly the polarity of the opinion word. Opinion can either be positive or negative.

In this project, we have focused on Unsupervised and Supervised Methods for ABSA of the restaurant reviews dataset. Some Unsupervised and supervised methods are proposed, implemented and evaluated. We firstly extract the aspect terms in each sentence then find out their polarities then detect the categories of these sentences and the polarity of each sentence of the testset.

**1. INTRODUCTION**

**1.1 General Introduction to the Topic**

Sentiment Analysis has been used by a host of major brands and businesses worldwide and it increasingly viewed as a vital task from a commercial standpoint. The majority of current approaches, however, attempt to detect the overall polarity of a sentence, paragraph, or text span, regardless of the entities mentioned (e.g. restaurants) and their aspects (e.g. food, service).

The main goal of ABSA is to extract the aspect terms to which these opinions are being targeted to.

Eg. The food was good, but the ambiance was poor.

Here, we have 2 aspects “food”, and “ambiance”. Also, polarities associated with both aspects are “positive”, and “negative” respectively.

We will use 2 labelled datasets for this problem, and will find aspects, corresponding polarities, and visualize them.

**1.2 Organization**

**Impetus**is a products, services and SAAS company headquartered in Los Gatos, USA. They have numerous development centers in INDIA including the ones in Noida, Indore, Gurgaon, and Bengaluru, India.

The main aim of the 1600 employee big company is to focus on creating innovative ways and products for analyzing the data for Fortune 500 clients and help them gain important business insights across different products and branches.

They have developed software products like StreamAnalytics and expertise across the Information Management ecosystem offering products and services including the Hadoop, NewSQL, NoSQL, and MPP databases. They have dedicated teams for ML and innovative visualization.

**1.3 Area of Computer Science**

Aspect Based Sentiment Analysis (ABSA) falls under the Machine Learning Branch of Computer Science Applications.

Machine learning is the science of teaching computers the process of learning about the data without explicitly programming it. Advancements in the field of ML have enabled revolutionary products like the self-driving cars, practical speech recognition, effective web search, and a vastly improved understanding of the human genome. Machine learning is so pervasive today that you probably use it dozens of times a day without knowing it.

Many researchers also think it is the best way to make progress towards human-level AI. In this class, you will learn about the most effective machine learning techniques, and gain practice implementing them and getting them to work for yourself.

**1.4 Hardware and Software Requirements**

* R studio
* TensorFlow
* CoreNLP

**2. PROBLEM DEFINITION**

**2.1 Dataset**

The data set provided by SemEval is a subset of Ganu et at(2009). It is in the XML format, and has separate ﬁles for Laptop and Restaurant reviews. The training data contains about 500 reviews ,which is 1606 sentences for Restaurants. The image attached below shows a part the XML ﬁle of the Restaurant data.



Fig.1: Snipptet of the Restaurant dataset XML ﬁle

For reach sentence, we have a target attribute which lists the aspect term, and a corresponding polarity attribute. The distribution of positive and negative sentiments in both the dataset is provided below:

|  |  |  |
| --- | --- | --- |
| Domain | Positive | Negative |
| Restaurant | 1198 | 408 |

**3. OBJECTIVE(S)**

Datasets consisting of customer reviews with human-authored annotations identifying the mentioned aspects of the target entities and the sentiment polarity of each aspect are given.

The goal is to identify the aspects of given target entities and the sentiment expressed towards each aspect. Eg. The food was good, but the ambiance was poor.

**4**. **BACKGROUND**

Various approaches have been adopted to identify aspects from sentences. Bing Lui et al. used frequency of noun phrases, followed by a redundancy pruning to identity the feature corresponding to a review. Yejin Choi et al. performed semantic tagging using conditional random ﬁelds with features based on Capitalization, syntactic chunking to extract sources of opinions from texts.

The best performing one uses a Conditional Random Field with features extracted using named entity recognition, POS tagging and parsing.

We try to augment this approach by using features not only based upon text processing techniques, but also on vector embedding of words and sentences. The motivation behind this being that the number of candidate aspect words of restaurant domain is rather restrictive. The task of polarity detection was addressed using various classiﬁcation techniques like Naive Bayes, SVM etc.

**5. METHODOLOGY**

In particular, the task consists of the following subtasks:

**5.1 Aspect Term Extraction**

Given a set of sentences with pre-identified entities (e.g., restaurants), identify the aspect terms present in the sentence and return a list containing all the distinct aspect terms. An aspect term names a particular aspect of the target entity.

For example, "I liked the service and the staff, but not the food”, “The food was nothing much, but I loved the staff”. Multi-word aspect terms (e.g., “hard disk”) should be treated as single terms (e.g., in “The hard disk is very noisy” the only aspect term is “hard disk”).

**5.2 Aspect Term Polarity**

For a given set of aspect terms within a sentence, determine whether the polarity of each aspect term is positive, negative, neutral or conflict (i.e., both positive and negative).

For example:

“I loved their fajitas” → {fajitas: positive}

“I hated their fajitas, but their salads were great” → {fajitas: negative, salads: positive}

“The fajitas are their first plate” → {fajitas: neutral}

“The fajitas were great to taste, but not to see” → {fajitas: conflict}

**5.3 Aspect Category Detection**

Given a predefined set of aspect categories (e.g., price, food), identify the aspect categories discussed in a given sentence. Aspect categories are typically coarser than the aspect terms of Subtask 1, and they do not necessarily occur as terms in the given sentence.

For example, given the set of aspect categories {food, service, price, ambience, anecdotes/miscellaneous}:

“The restaurant was expensive, but the menu was great” → {price, food}

**5.4 Aspect Category Polarity**

Given a set of pre-identified aspect categories (e.g., {food, price}), determine the polarity (positive, negative, neutral or conflict) of each aspect category.

For example:

“The restaurant was expensive, but the menu was great” → {price: negative, food: positive}

**6. IMPLEMENTATION DETAILS**

**6.1 Unsupervised Approach**

We load the XML file into our R studio environment and extract all the text from it.

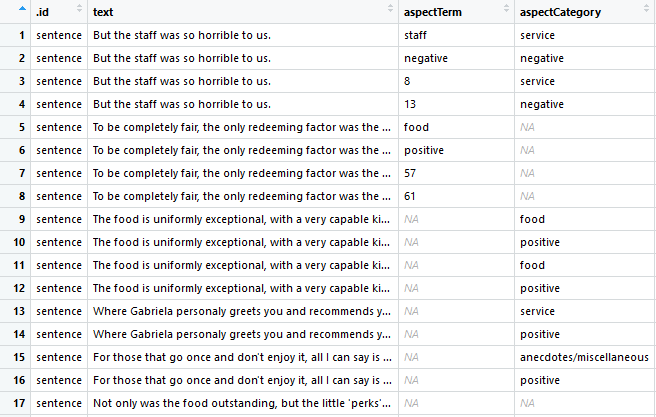


Fig.2:Docxml Dataframe Snapshot

We now need to extract terms from our sentence. We need to build a custom built function for extraction and cleansing of the data. Parts-of-Speech Tagging tags each word in different class groups like “Nouns”, ”Prepositions” and “Adjectives” etc.

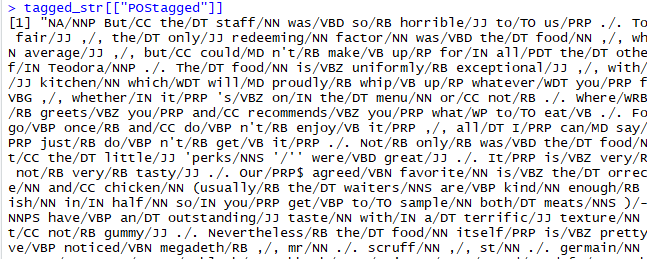


Fig.3: POS Tagged result

We now separate the aspects by filtering our POS tagged words for “/NN\*”.We then remove the “/NN\*” part of the aspect. Using this we extract all the aspects from the testset.

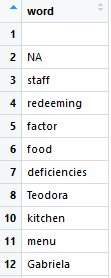


Fig.4: Aspects Dataframe Snippet

We now find the aspects in each sentence so as to link them later to make governor-Dependent pairs.

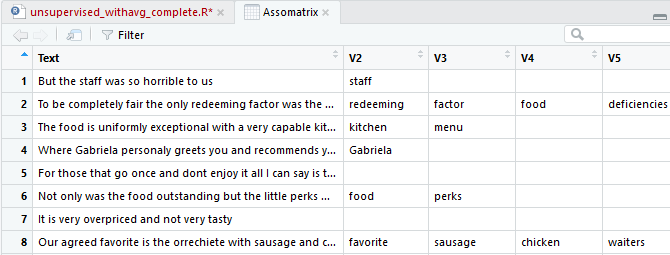


Fig.5: Sentence – Aspect Matrix

We have taken special care of improving the efficiency of the program by using unique terms after stemming and lemmatization.

We get a highly sparse matrix when we make a Document-Term matrix for our aspects and text data. This is a necessary perquisite for passing as argument to the Apriori().

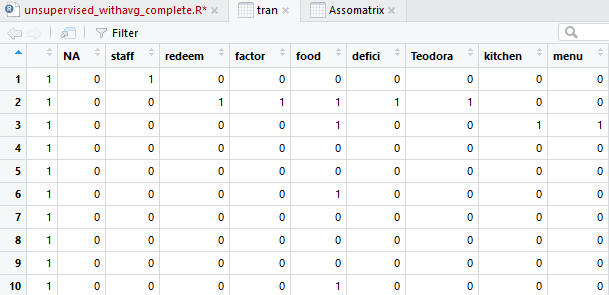


Fig.6: Document-Term Matrix

We use Apriori’s algorithm to find the most frequent itemsets and remove the unnecessary aspects by setting a support and confidence threshold.

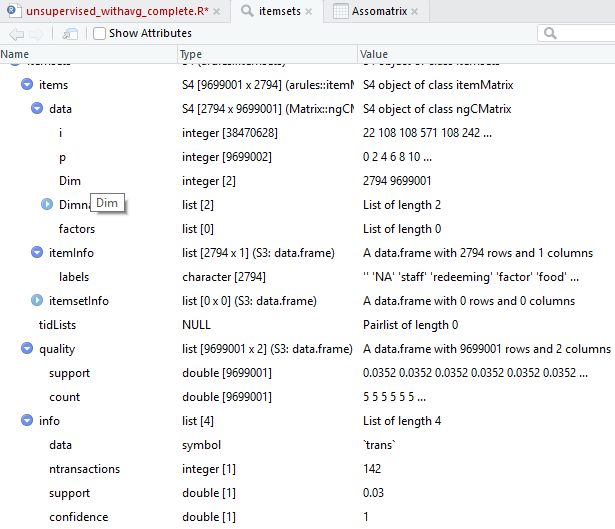


Fig.7: Apriori Algorithm

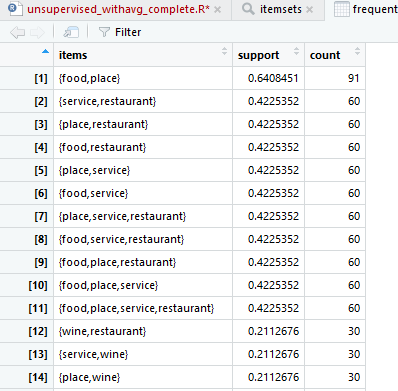


Fig.8: Most Frequent Itemsets

Inference: {food,place} is the most frequent itemset which is present in our data set.

It is present 91 times in 1000 sentences.

We now find Governor – Dependent pairs using the Stanford Dependency parser (SDP).

GD pairs help us find relationship between 2 words in a sentence.

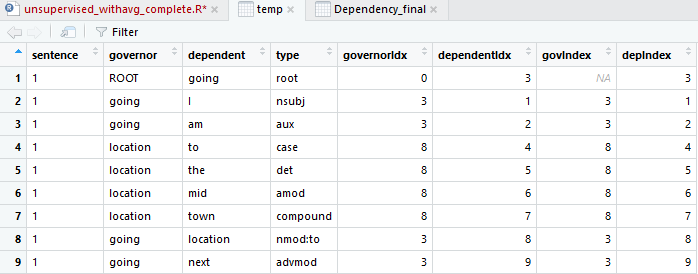


Fig.9: Governor-Dependent Pairs

We also need to consider negated words like” did not like the food”.

By default like has a polarity of +1 but in this case the sentence will have -1 since like is negated by the use of “not”

We need to reverse the polarity of those words which are paired with negation words.

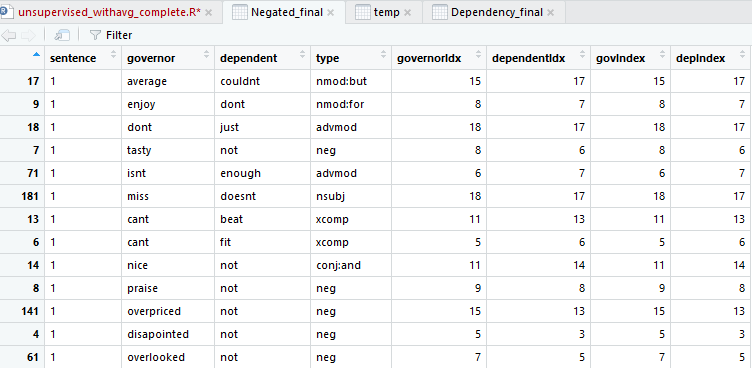


Fig.10: Negation list Filtered

We then create an extra column in the Dependency\_final dataframe for the polarity of the GD pairs.

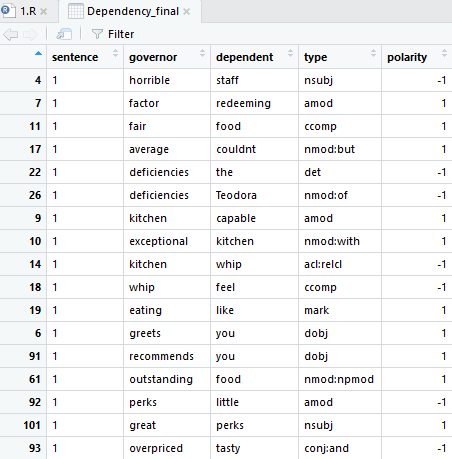


Fig.11: Polarity for GD Pairs

Since we now have GD pairs with polarity and the negated pairs, we can correctly predict the polarity of each sentence of the test set.

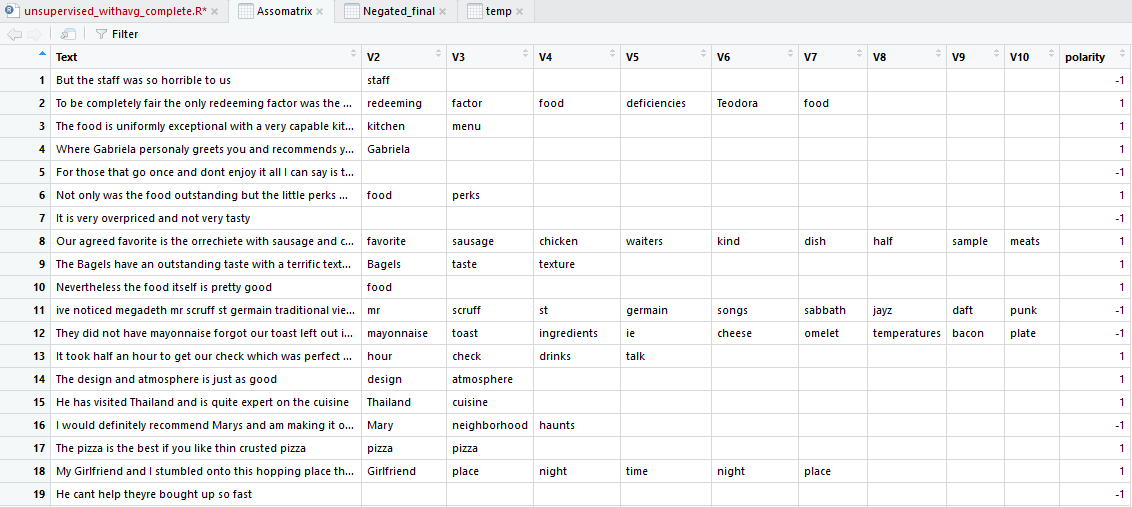


Fig.12: Polarity of each sentence

To find the accuracy of my model, we filter the testset to find the sentences for which we have the polarity.

Polarity in the test data could either be positive or negative. If the polarity is positive then we convert it to 1 and -1 if it is negative. We now match the results with the polarity of the test set.

1451 / 1892 polarities were predicted accurately which gives us an **Accuracy** of **0.766**.

**6.2 Supervised Approach**

We will now try to use a supervised approach and train a model to predict the polarity of the test data.

Supervised Approach has 4 Subtasks:

**6.2.1 Opinion Target Extraction**

The objective of OTE is to extract all opinion target expressions in a text review, OTE could be a word or multiple words.

For this purpose, we have used CRF (Conditional Random Field) which have proved its performance in information extraction.

Eg. In the context “food was good”, food is the Opinion Target Word.

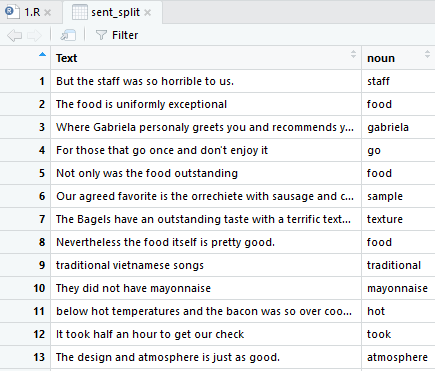


Fig.13: Context – Opinion Target word Extraction

We find these OTE’s using the same approach as we used in our unsupervised approach.

We used the same POS Tagging method and filtering that we had used earlier.



Fig.14:POS tagged words

**6.2.2 Feature Extraction (Sentiment Polarity)**

We use these features to refine our result and increase the accuracy of our model by providing it more data to learn from.

**6.2.2.1 Word N-Grams**

We find Unigrams and Bigrams for each context.

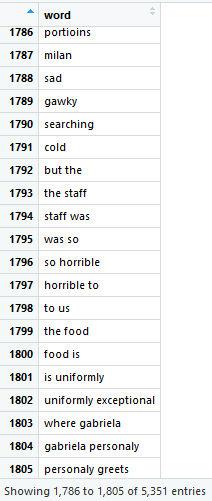


Fig.15:N-Gram DataFrame

**6.2.2.2 Document Term Frequency Extraction**

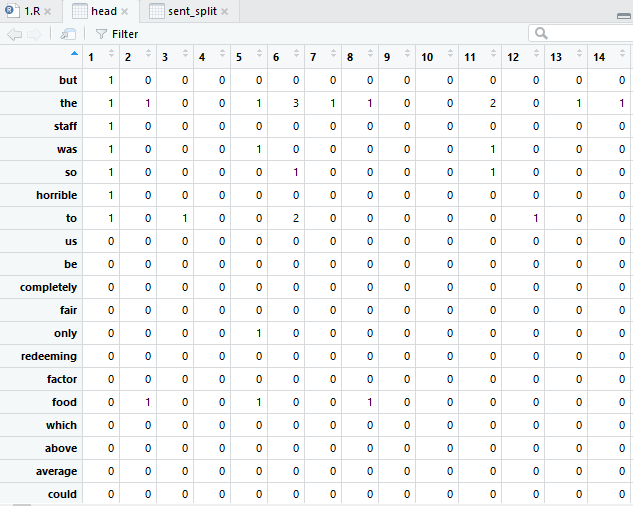


Fig.16:Term-Document Frequency Matrix

The Term-Document Frequency helps the model to calculate values like TD – IDF and TF-IDF .

**6.2.2.3 Z Score Features**

Z score can distinguish the importance of each term in each class. Thus, Z score can be seen as a standardization of the term frequency. We compute the Zscore for each term ti in a class Cj (tij) by calculating its term relative frequency tfrij in a particular class Cj , as well as the mean (mean i)which is the term probability over the whole corpus multiplied by nj the number of terms in the class Cj, and standard deviation (sdi) of term ti according to the underlying corpus (see Eq.1).

Zscore(ti) = (tfrij –mean i)/sdi (1)

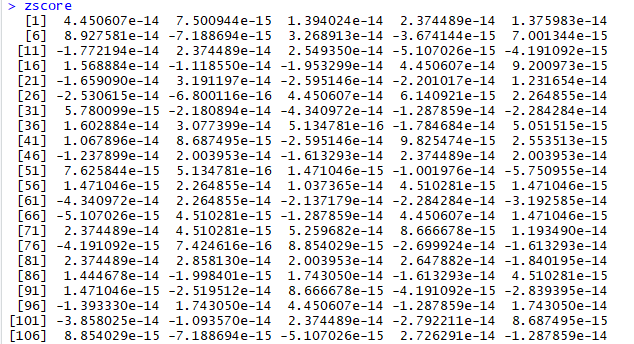


Fig.17: Z-Score for each context

We find the Z-score of each context by adding the zscore of all its words.

**6.2.2.4 Sentiment Lexicon Based Features**

For each context the number of positive words, the number of negative ones, the number of positive words divided by the number of negative ones and the polarity of the last word are extracted from manual constructed lexicons.

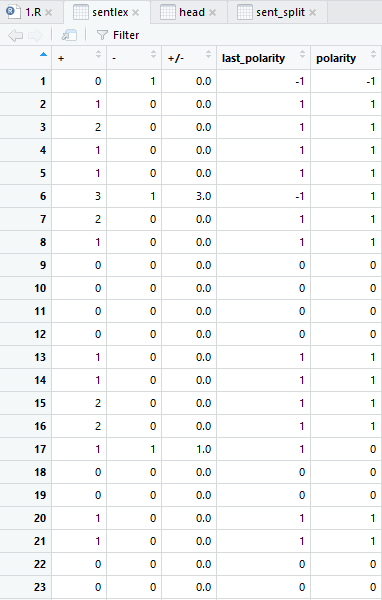
****

Fig.18:Sentlex Dataframe Snapshot

**6.2.3 Model Generation**

We divide our dataset into 2 parts: training data (contains 400 reviews) and testing data (contains 100 reviews).

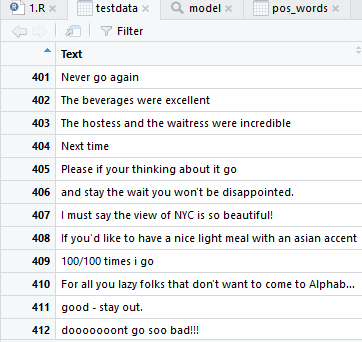


Fig.19: Testdata Dataframe Snippet

Before passing the datasets into the model() and predict(), we need to process it by converting it to plaintext format after vectorising the data and then converting it to a corpus.

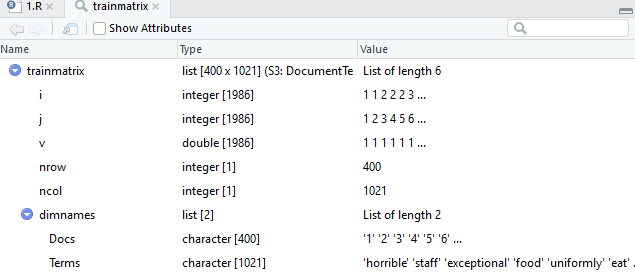


Fig.20: Trainmatrix snippet

We have used the Naïve Bayes Training model using the trainmatrix data and by selecting the traindata$polarity class vector.

By selecting the polarity vector, we set the prediction variable using the polarity column of the traindata dataset. The polarity column contains the correct polarity for each context and gives our Naïve Bayes model a vector to model the other features around.

The Naïve Bayes will classify the testdata using the all these features and then set the polarity in the testdata$polarity column. It will classify a context (sentence) based on the similarity of the other features like Z-score or TF-IDF values.

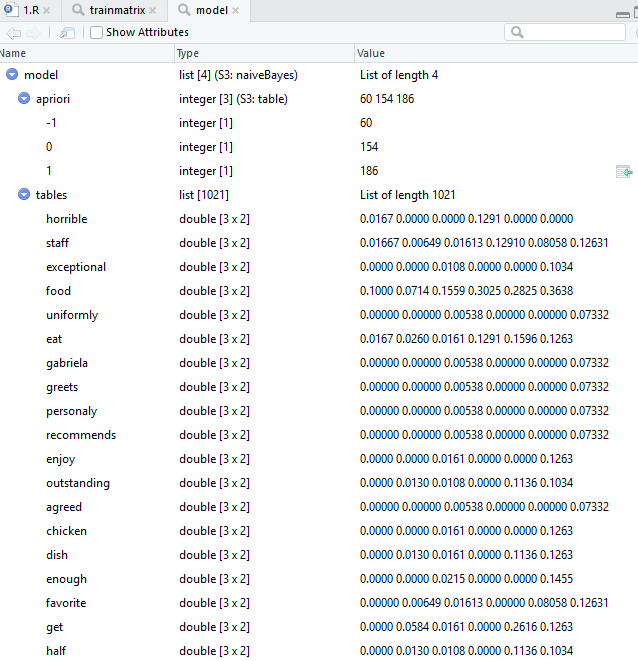


Fig.21: Dataframe snapshot of the Model

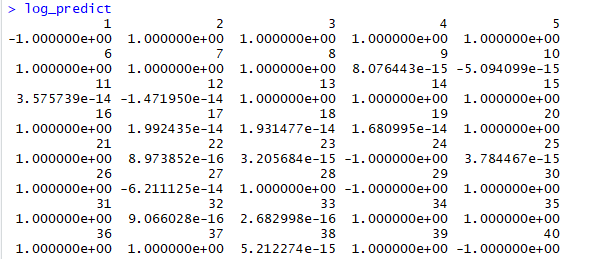


Fig.22: Results of predict()

The log\_predict function returns -1 or 1 for positive and negative polarity and -Inf<x<-1 or 1<y<Inf for incorrect responses.

We get an **Accuracy** of **.81** which we found out by comparing the results of the testing data with predefined polarities of the provided data.

**7. PROJECT SCHEDULE**

* *January 2018*
* Study Part of speech Tagging in NLP
* *February 2018*
* Implementation of unsupervised ABSA NLP program
* *March 2018*
* Implementation of supervised ABSA NLP program
* Documentation
* *April 2018*
* Submission of report & evaluation
* Study DeepLearning Technology and begin implementation
* *May 2018*
* ABSA implementation using a DeepLearning Approach ( TensorFlow)
* *June 2018*
* Submission of report & evaluation

**8. REFERENCES**

* M. Hu and B. Liu, “Mining and summarizing customer reviews”. Proceedings of the 10th KDD, pp. 168–177, Seattle, WA, 2004.
* Ankit Singh, Enayat Ullah , “Aspect based Sentiment Analysis”. DOI: <https://cse.iitk.ac.in/users/cs365/2015/_submissions/enayat/report.pdf>
* Semeval 2014. “SemEval-2014 Task 4”. In: (2014). DOI: <http://alt.qcri.org/semeval2014/task4/>
* Hussam Hamdan,Patrice Bellot and Frederic Bechet ,”Lsislif: CRF and Logistic Regression for Opinion Target Extraction and Sentiment Polarity Analysis“. DOI: <http://anthology.aclweb.org/S/S15/S15-2128.pdf>
* Samuel Brody and Noemie Elhadad,” An Unsupervised Aspect-Sentiment Model for Online Reviews”. DOI: <http://people.dbmi.columbia.edu/noemie/papers/naacl10.pdf>

PROJECT DETAILS

|  |  |  |  |
| --- | --- | --- | --- |
| *Student Details* | | | |
| **Student Name** | **SHASHWAT GUPTA** | | |
| Register Number | 140905047 | Section / Roll No | C - 06 |
| Email Address | [Shashwat2211@gmail.com](mailto:Shashwat2211@gmail.com) | Phone No (M) | 9620193750 |
|  | | | |
| *Project Details* | | | |
| **Project Title** | Aspect based Sentiment Analysis using NLP (R) and Deep Learning | | |
| Project Duration | 5 Months | Date of reporting | 22 - Jan - 2018 |
|  |  | | |
| *Organization Details* | | | |
| **Organization Name** | **Impetus Infotech India Pvt. Ltd.** | | |
| Full postal address with pin code | SDF No. K-13 to 16, NSEZ, Phase-II, NOIDA - 201305 (U.P.) India | | |
| Website address | Impetus.com | | |
|  |  | | |
| *External Guide Details* | | | |
| **Name of the Guide** | **Mr. Vivek Gupta** | | |
| Designation | Senior Analytics Engineer | | |
| Full contact address with pin code | Data Science Practice , SDF No. K-13 to 16, NSEZ, Phase-II, NOIDA - 201305 (U.P.) India | | |
| Email address | [Vivek.gupta@impetus.co.in](mailto:Vivek.gupta@impetus.co.in) | Phone No (M) | 9910464101 |
|  |  | | |
| *Internal Guide Details* | | | |
| **Faculty Name** | **Mrs. Archana Praveen Kumar** | | |
| Full contact address with pin code | Dept of Computer Science & Engg, Manipal Institute of Technology, Manipal – 576 104 (Karnataka State), INDIA | | |
| Email address | [Archana.kumar@manipal.edu](mailto:Archana.kumar@manipal.edu) | | |