"CAPTURING THE STARS" – SENTIMENT ANALYSIS AND SUMMARIZATION OF ONLINE RESTAURANT REVIEWS

Project report submitted in fulfillment of the requirement for the degree of Bachelor of Technology

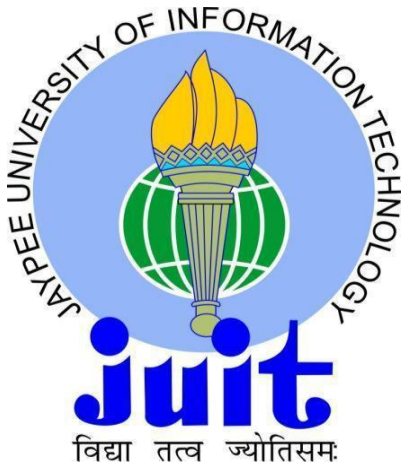
In

Computer Science and Engineering/Information Technology

By

Ayushi Singh (141344)

Under the supervision of
Mr. Amol Vasudeva
to



Department of Computer Science & Engineering and Information Technology
Jaypee University of Information Technology Waknaghat, Solan 173234, Himachal Pradesh

CERTIFICATE

Candidate's Declaration

I hereby declare that the work presented in this report entitled **"Capturing the Stars- Sentiment Analysis and Summarization of Online Restaurant Reviews"** in complete fulfillment of the requirements for the award of the degree of **Bachelor of Technology** in **Computer Science and Engineering/Information Technology** submitted in the department of Computer Science & Engineering and Information Technology**,** Jaypee University of Information Technology Waknaghat is an authentic record of my own work carried out over a period from August 2017 to May 2018 under the supervision of **Mr. Amol Vasudeva,** Assistant Professor in Computer Science & Engineering and Information Technology Department.

The matter embodied in the report has not been submitted for the award of any other degree or diploma.

Ayushi Singh, 141344

This is to certify that the above statement made by the candidate is true to the best of my knowledge.

Mr. Amol Vasudeva
Assistant Professor
Computer Science & Engineering and Information Technology Department
Dated: 10 May, 2018

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to my supervisor **Mr. Amol Vasudeva** for providing his invaluable guidance, comments and suggestions throughout the course of the project. He took keen interest and guided me all along in my project work titled — **Capturing the Stars- Sentiment Analysis and Summarization of Online Restaurant Reviews**, till the completion of my project by providing all the necessary information for developing the project.

Being grateful for his support, constructive criticism and valuable suggestion throughout the project work, I'm sincerely very thankful for all his efforts.

Last but not the least, I thank TripAdvisor's authority for allowing me to fetch and use data from their website.

TABLE OF CONTENTS

LIST OF ABBRVIATIONS

SVM…….Support Vector Machines

NYBS……...Naïve Bayes

ML……Machine Learning

NLP……..Natural Language Processing

ANN…….Artificial Neural Network

BI…….Business Intelligence

MaxEnt……Maximum Entropy

API……Application Program Interface

LIST OF FIGURES

Title **Page No.**

LIST OF GRAPHS

Title **Page No.**

LIST OF TABLES

Title **Page No.**

ABSTRACT

In this model we use supervised machine learning approach where features like word unigrams and frequency counts are used to train classification and regression models. We have used sentiment analysis as a hybrid of lexicon based and machine learning algorithm. Aspects used for rating are: food, service, atmosphere, value. We have used reviews from TripAdvisor of about 30 restaurants with an average of 400 reviews per restaurant containing around eight sentences per review. In our experiments, we considered aspect rating predictors as independent from each other. For each rated aspect, predictor models were trained independently, by the end we observed individually aspect ratings are highly correlated with overall ratings.

With the help of Data Cleaning and Text Extractor we broke review texts into a set of text fragments that are of use and formed a summary. For certain experiments, we used only the unigram words for analyses. Overall ratings given by reviewers can be biased, so we did not use them directly as input. Although we have trained the predictors using aspect rating provided by reviewers, our model is able to predict aspect ratings that depend only on the input text and not on reviewer's biasness. We found that only 62% of user given ratings have supporting text for ratings of these aspects in the reviews. Besides unigrams in the review text, we also used the counts of positive and negative polarity words and their differences. Polarity labels are obtained from a dictionary of about 700 words. This dictionary was created by first collecting words used as adjectives in a corpus of un–related review text.

We then retained only those words in the dictionary that, in a context free manner generally conveyed positive or negative evaluation of any object, event or situation. We have used Machine learning to predict ordinal ratings from textual data and used MaxEnt classification to minimize classification error. Then we partitioned the data into 90% for training and 10% for testing. Random distribution ensures training and test data are identical. All the results are averages of 10–fold cross– validation over 12,000 review examples.

CHAPTER 1: PROJECT OBJECTIVE

1.1 INTRODUCTION

Natural Language Processing:

"Natural Language Processing is a field that does understanding and manipulation of human language into a language understandable by computer, and it's blooming with possibilities for news-gathering". "You usually hear about it in the context of analyzing large pools of legislation or other document sets, attempting to discover patterns or root out corruption." With help of NLP computers analyze, understand, and derive meaning from human language into a smart and useful way. With the help of NLP, developers have organized and structured the knowledge to implement tasks of 'automatic summarization. Those are: following: named entity recognition, and sentiment analysis. Others are: translation, and topic segmentation. Some more are: relationship extraction, and speech recognition.'

"Apart from common word processor operations that have been treating text like a sequence of symbols, natural language processing takes into consideration the hierarchical structure of this language, that means, many different words combine to form a phrase, and several phrases combine to form a sentences and, finally, these are the sentences that convey ideas," By analyzing language for its meaning, NLP systems have been performing the tasks of converting speech to text and then automatically translating between languages and correcting grammar alongside for a very long time. Natural Language Processing is used to analyze text, thereby allowing machines to understand how human speak.

This human-computer interaction promotes processing of real-world applications like: sentiment analysis and automatic text summarization. Some other tasks like named entity recognition and topic extraction. Also the tasks like relationship extraction and parts-of-speech tagging, with stemming, and many more. NLP is most commonly used for machine translation, text mining and for automated question answering. In computer science Natural Language Processing is characterized as a hard problem. A common characteristic with all human languages are; they are rarely precise, or plainly spoken. To understand human language is to understand not only the words, but the concepts and also how they're linked together to create some meaning. Although language is one of the easiest things for human-beings to learn, it is because of the

ambiguity of languages that makes natural language processing a difficult problem for even computers to master.

NLP algorithms are generally based on machine learning algorithms. Instead of manually coding large sets of rules, NLP very much depends on ML to learn these rules automatically and analyze the given set of examples (i.e. a large corpus, like a book, small as collection of sentences), and making a statistical inference. Generally, more the data analyzed, more accurate the model will be.

- **Summarize blocks of text** using Summarizer extracting the most important and central ideas and ignoring irrelevant information.
- Create a **chat bot** using Parsey McParseface, a language parsing deep learning model made by Google that uses Point-of-Speech tagging.
- **Keyword tags are generated automatically** from the content with the help of AutoTag, that leverages LDA, it is a technique that helps in discovering those topics that are contained within a body of the given text.
- **Identify the type of entity extracted**, for example it being a person, place, or organization using Named Entity Recognition.
- Use Sentiment Analysis for **identifying the sentiment of a string of text**, from negative to neutral to positive.
- **Reduce words to their root**, or stem, using Porter Stemmer, or **break up text into tokens** by tokenization using Tokenizer.

Open Source NLP Libraries will provide the algorithmic building blocks of NLP in real-world applications. Since this algorithm provides a free APIs with end-points for many of these algorithms, so there is no need to develop separate infrastructure and servers.

Machine Learning:

Machine learning is a subset of artificial intelligence (AI). Generally machine learning's goal is to understand the structure of data and fitting that data into models that can be understood and efficiently utilized by different people.

Even though machine learning is a field within computer science, it is different from traditional computational approaches. In traditional computing, algorithms are set of explicitly programmed instructions used by computers for calculation and problem solving. Instead Machine learning algorithms allow computers to train on data-inputs and use statistical-analysis to order data-output values that fall within an anticipated specific range. It's because of this that machine learning helps computers in building models from sampled data for automating decision-making processes based on data-inputs.

Any technology user today has been benefitted from machine learning. The technology of Facial recognition allows social media websites and platforms in helping users tag and share photos of friends. The technology of Optical character recognition (OCR) converts images of text into movable type. Recommendation engines, which are powered by machine learning, suggest what movies or television shows to watch next based on each user's preferences. Self-driving cars that rely highly on machine learning to navigate, they will soon be available to consumers.

Machine learning is a continuously developing field. It is because of this, there are some important considerations that should be kept in mind as you work with machine learning methodologies, or if you are analyzing the impact of machine learning processes.

Some common machine learning methods of supervised and unsupervised learning are, common algorithmic approaches in machine learning, including the k-nearest neighbor algorithm, decision tree learning, and deep learning. We are going to explore which programming languages are most frequently used in machine learning, and providing you with the positive and negative attributes of each language. We will discuss biases that are caused by machine learning algorithms, and considerations that should be kept in mind to prevent these human-biases while building algorithms.

Machine Learning Methods

In machine learning, certain tasks are classified into some broad categories. These categories are based on how learning is received by the system and how feedback on the learning is given to the system that has been developed.

Two of the most widely adopted machine learning methods is **supervised learning** which is used for training algorithms based on examples of input, output data that is labeled by human-beings, and **unsupervised learning** which provides the algorithm with no labeled data in order to allow it to find structure within its input data. We'll explore these methods in detail.

Supervised Learning

For supervised learning, the computer is provided with the example of inputs that are labeled with their desired outputs. The purpose of this method is that the algorithm should be able to "learn" by comparing its actual output with the "taught" outputs to find errors, and then modify the model in the required way. Supervised learning uses patterns to predict label values on the additional unlabeled data.

For example, with supervised learning, an algorithm may be fed data with images of sharks labeled as `fishes` and images of rivers labeled as `water`. By being trained on this data, the supervised learning algorithm should be able to later identify unlabeled shark images as `fish` and unlabeled river images as `water`.

A very common use of supervised learning is using historical data for predicting statistically likely future events. It may use historical stock market information to anticipate upcoming fluctuations, or it can be employed to filter out spam emails. In supervised learning, tagged photos of clouds can be used as input data to classify untagged photos of clouds.

Unsupervised Learning

For unsupervised learning, data is unlabeled, so the learning algorithm has to find commonalities in its input data. As we all know unlabeled data are more abundant than labeled data, machine learning methods that facilitate unsupervised learning are particularly very important.

The final aim of unsupervised learning may be very straight-forward like discovering hidden patterns within a data-set, alongside it may also have a goal of feature-learning, which allows the computational-machine to automatically find the representations that are required to classify raw-data.

Unsupervised learning is generally used for transactional data. Suppose you may have a large dataset of customers and their purchases, but as a human-being you will likely not be able to make sense of what similar attributes can be drawn from customer profiles and their types of purchases. With this data fed into an unsupervised learning algorithm, it may be determined that women of a certain age range who buy unscented soaps are likely to be pregnant, and therefore a marketing campaign related to pregnancy and baby products can be targeted to this audience in order to increase their number of purchases.

Without knowing the "correct" answer, unsupervised learning methods can look at complex data that is more expansive and seemingly unrelated, in order to organize it in a potentially meaningful way. Unsupervised learning is majorly used for anomaly detection including for forged credit card purchases, and recommender systems that recommend what products you may like to buy. With unsupervised learning, untagged photographs of clouds is used as input data for the algorithm, for finding likenesses and classifying cloud photographs together.

Without knowing the "correct" answer, unsupervised learning methods can look at complex data that is more expansive and seemingly unrelated, in order to organize it in a potentially meaningful way. Unsupervised learning is majorly used for anomaly detection including for forged credit card purchases, and recommender systems that recommend what products you may like to buy. With unsupervised learning, untagged photographs of clouds is used as input data for the algorithm, for finding likenesses and classifying cloud photographs together.

Approaches

Machine learning is very much related to computational-statistics, so having background knowledge in statistics will be very much useful for leveraging and understanding machine learning algorithms.

People who not have studied statistics, for them it can be helpful to first define regression and co-rrelation, since they are commonly used techniques for investigating the relationship between quantitative variables. **Regression** is used to examine the relationship between one dependent and another independent variable, since regression statistics is used to predict the dependent variable when the independent variable is known, regression technique enables prediction capabilities. **Correlation** is a measure of association between two variables that are not designated as either independent or dependent.

Approaches to machine learning are continuously being developed. Here, we are going to go through few of the popular approaches that are being used in machine learning these days.

k-nearest neighbor

The k-nearest neighbor algorithm is a pattern recognition model that can be used for regression and classification. Often abbreviated as k-NN, the **k** in k-nearest neighbor stands for a positive integer, which is generally very small. In either regression or classification, the input will consist of the k closest training examples within a space.

Here we focus on k-NN classification. Note that the output is class-membership in this method. Every time this will assign a new object to the class that is most common among its k nearest neighbors. In the case of k = 1, the object is assigned to the class of the single nearest neighbor.

Now look at the example of k-nearest neighbor. In the below diagram, there are blue diamonds and orange stars. These belong to two separate classes of objects: the diamond class and the star class.

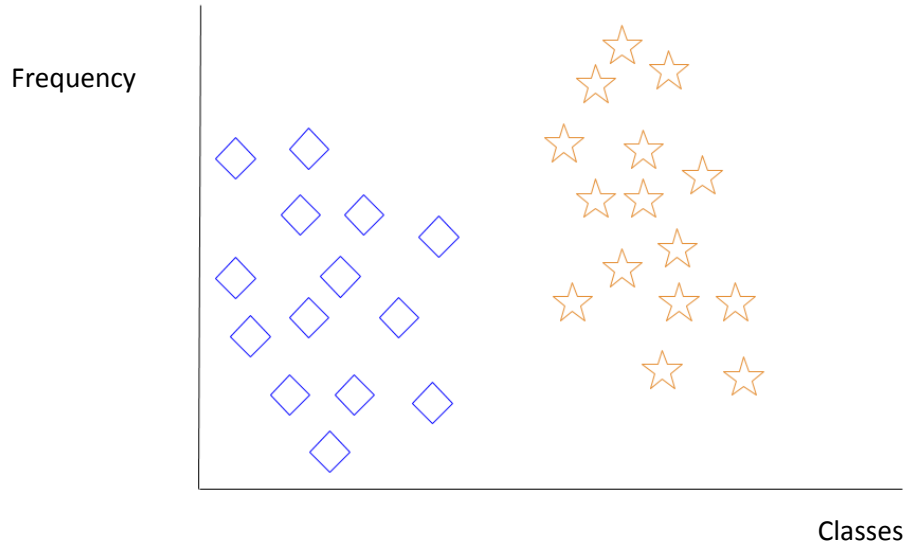


Figure 1 : Two separate classes: diamond class and star class

When a new object is added to the space - in this case a green heart - we will want the machine learning algorithm to classify the heart to a certain class.

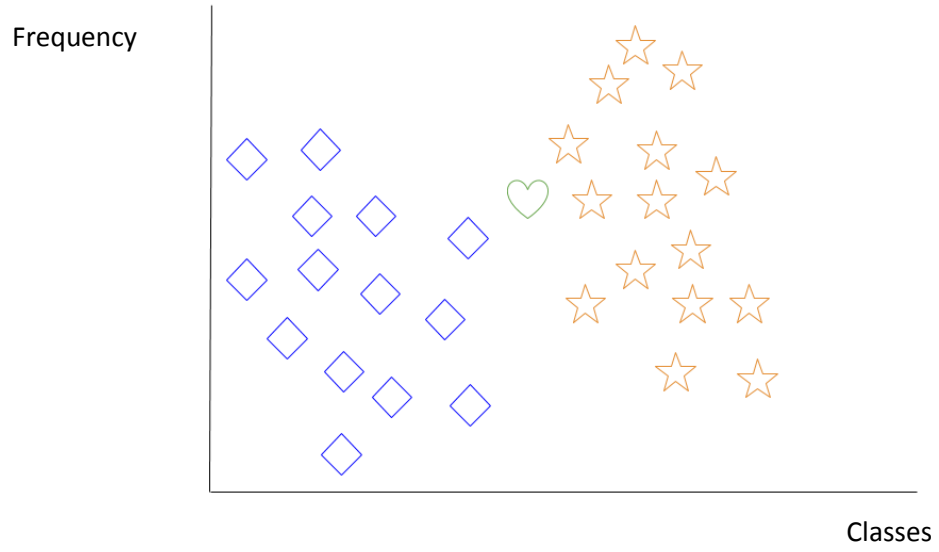


Figure 2 : New object inserted

When we choose k = 3, the algorithm will find the three nearest neighbors of the green heart and classify it to either the diamond class or the star class.

In this diagram, the three nearest neighbors of the green heart are one diamond and two stars. Therefore, the algorithm will classify the heart with the star class.

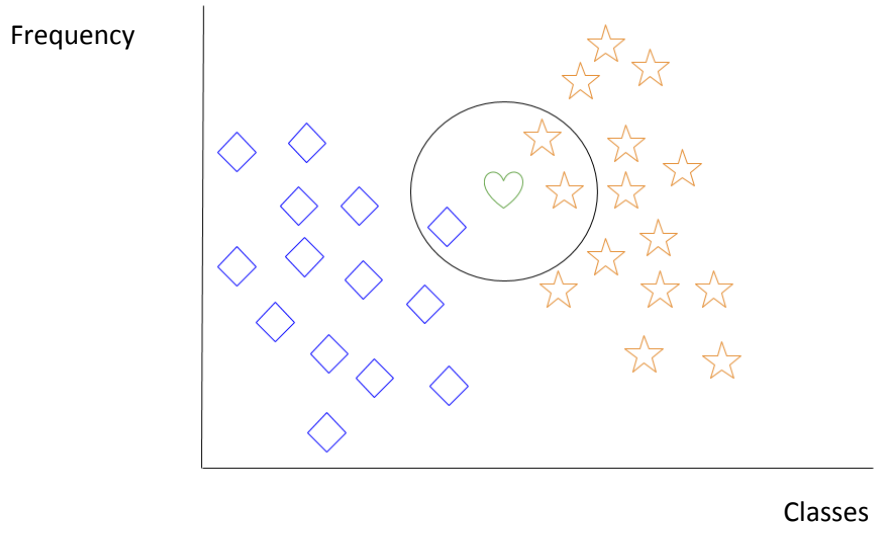


Figure 3 : Algorithm will classify the heart with the star class

Among all the machine learning algorithms, k-nearest neighbor is considered to be a type of easiest or of "lazy learning" since generalization beyond the training data does not occur until a query is made to the system in use.

Decision Tree Learning

Generally, decision trees are employed for visual representation of decisions and show informed decision making. When working with data mining and machine learning, decision trees are used as a predictive model. These models are used to map the observations about data to the conclusions about the data's target value.

The goal of decision tree learning technique is to create a model that will predict the value of a target based on input variables. In the predictive model, the data's attributes that are determined through observation are represented by the branches, and the conclusions about the data's target value are represented in the leaves. When "learning" a tree, the source data is divided into subsets based on an attribute value test, same method is repeated on each of the derived subsets recursively. Once the subset at a node has the equivalent value as its target value has, the recursion process will be complete. Let's look at an example of various conditions that can determine whether or not someone should go trekking. This includes barometric pressure conditions as well as weather conditions.

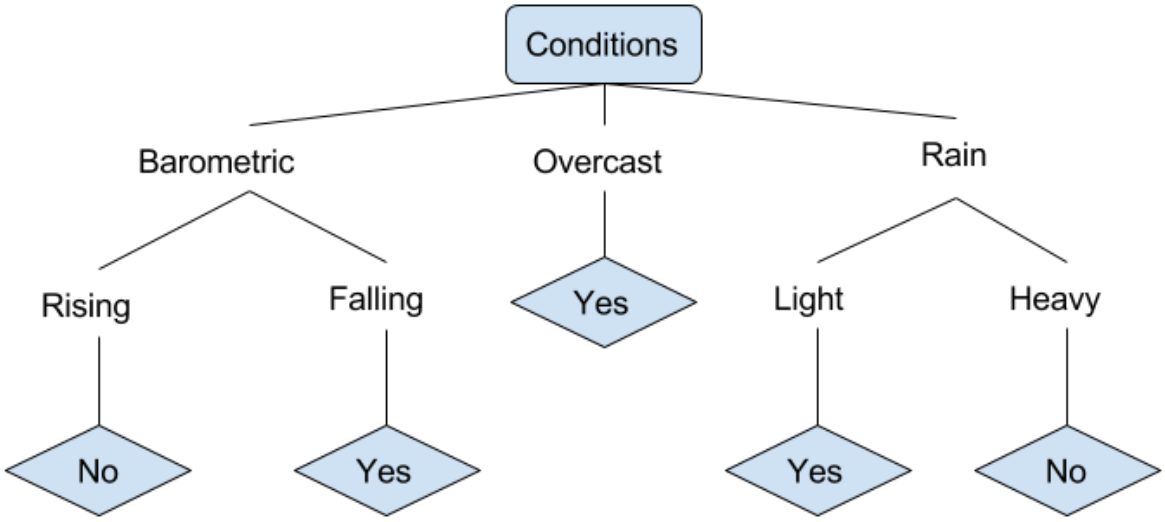


Figure 4 : Decision Tree [8]

In the above simplified decision tree, an example is classified by sorting it through the tree to the respective leaf nodes. This then returns the classification associated with the particular leaf, which in every case is either a 'Yes' or a 'No'. The tree classifies a day's conditions based on whether or not it is suitable for going trekking.

Real life classification tree data set would have a many more features than what are outlined in the above example, but relationships have to be straightforward to anticipate. When working with decision tree learning, several determinations need to be made, including what features to choose, what conditions to use for splitting and understanding when the decision tree has reached its very end.

Deep Learning

Deep learning imitates how the human-being brain can process light and sound stimuli into vision and hearing. Deep learning architecture is inspired by biological neural networks and similarly consists of multiple layers in an artificial neural network made up of hardware and GPUs.

In order to extract or transform features (or representations) of the data, deep learning uses a cascade of nonlinear processing unit layers. The output of a single layer serves as the input of the successive layer. In deep learning technique, algorithms can be either supervised and serve to classify data, or unsupervised and perform pattern analysis.

Among the entire machine learning algorithms that are currently being used and developed, deep learning uses the maximum data and has been able to over-power human-beings in some cognitive tasks, because of these attributes, in the field of artificial intelligence dep learning has become very important.

Speech recognition and Computer vision both realized significant advances from deep learning approaches. IBM Watson is a well-known example of a system that implements deep learning technique.

Programming Languages

One considers the skills listed on current job advertisements, when choosing a language to specialize in with machine learning. They also see the libraries that are available in various languages and the ones that are used for machine learning processes.

Python has been the most used programming language in the machine-learning professional field. Python is followed by Java, then R, then C++ (this information has been taken from indeed.com, December 2016).

Python- It is popular because there have been increase in the development of deep learning frameworks available for this language recently, including TensorFlow, PyTorch, and Keras. As a language python readable syntax and its ability to be used as a scripting language, it also proves to be powerful and straightforward both for working with data directly and preprocessing the data. The scikit-learn machine learning library is built on top of several existing Python packages that Python developers are mostly familiar with, namely NumPy, SciPy, and Matplotlib.

Java is widely used in enterprise programming, and is generally used by front-end developers of desktop application, who are also working on machine learning at the enterprise level. Usually java is not the first choice for people who are new to programming and who want to learn about machine learning, but it is definitely favored by those with a background in Java development, and want to apply to machine learning. In industry machine learning applications, Java is used more than Python for fraud detection use cases, network security, and cyber-attack.

Machine learning libraries for Java are Deeplearning4j, an open-source and distributed deep-learning library written for both Java and Scala; MALLET (**MA**chine **L**earning for **L**anguag**E** **T**oolkit) allows for machine learning applications on text, including natural language processing, document classification, topic modeling, and clustering; and Weka, a collection of machine learning algorithms that is used for data mining tasks.

R is an open-source programming language used basically for statistical computing. It popularity has grown in recent years, and is used by many in academic field. R is not typically favoured in industry production environments, but has risen in industrial applications due to increased interest in data science. Popular packages for machine learning in R are:

- For creating predictive models, caret is used (**C**lassification **A**nd **RE**gression **T**raining).
- For classification and regression, random forest is used .
- For statistics and probability theory, e1071 is used, it contains functions.

Machine learning and artificial intelligence uses the C language in the applications such as gaming robotics which also includes the locomotion of robot. Embedded computing hardware developers and electronics engineers generally favor C++ or C in machine learning applications because of their proficiency and level of control in the language. Some machine learning libraries you can use with C++ include the scalable mlpack, Dlib offering wide-ranging machine learning algorithms, and the modular and open-source Shark.

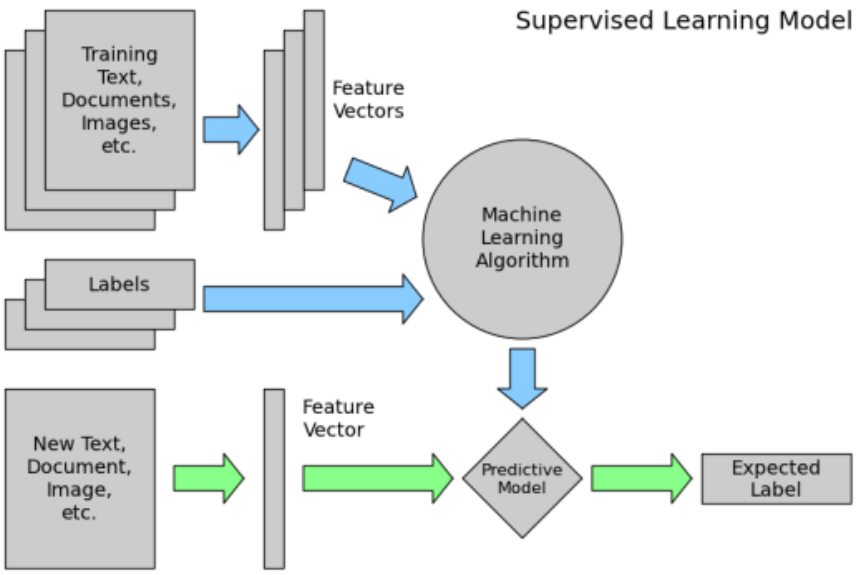


Figure 5 : Supervised Learning Model [9]

Human Biases

Even though data and computational-analysis makes us think that we are receiving objective information, this is not the case; being based on data does not mean that machine-learning outputs are neutral. Human-being bias plays an important role in how data is collected, way of organizing, and finally in the algorithms that determine how machine learning is going to interact with that input data.

Suppose people are providing images for "bird" as data to train an algorithm, and these people overwhelmingly select images of eagles and vultures, a computer may not classify a sparrow as a bird. This would create a bias against sparrow as birds, and sparrows would not be counted as birds.

Some AI and machine learning programs exhibit human-like biases that include race and gender prejudices.

Since human-being bias can negatively impact others, it is very important to be aware of it, and to also work towards eliminating it to the maximum level. It should be ensured that diverse people are working on a project and that diverse people are testing and reviewing it. Many others have called for regulatory third parties to monitor and audit algorithms, building alternative systems that can detect biases, and ethics reviews as part of data science project planning.

Conclusion

Above details reviewed some of the uses of machine learning. Some common methods and popular approaches used in the field, and suitable machine learning programming languages, also covered some issues to keep in mind in terms of unconscious biases being replicated in algorithms.

Opinion Mining:

It is a hard challenge for language technologies. The task of automatically classifying a text written in a natural language into a positive or negative feeling, opinion or, is sometimes so complicated.. Everyone's personal interpretation is different from others, it is also affected by cultural factors and each person's experiences. The shorter the text more difficult the task becomes, like in the case of messages on social networks like Facebook or Twitter.

Figure 6 : Sentiments

Two approaches

The problem has been tackled majorly from two different approaches:

- computational learning techniques and
- semantic approaches.

Semantic approaches is by the use of dictionaries of words (*lexicons*) with semantic orientation of polarity or opinion. Systems generally preprocess the text and divide it into words, with proper removal of stop words and a linguistic normalization with stemming or lemmatization, and then check the presence or absence of each term of the lexicon, using the sum of the polarity values of the terms for assigning the global polarity value of the text. Systems also include:

- More or less advanced treatment of modifier terms (such as *very*, *too*, *little*) that increase or decrease the polarity of the accompanying terms and
- Negations or inversion terms (such as *no*, *never*), these words reverse the polarity of the terms to which they are related to.

Learning-based approaches consists on training a classifier using any supervised learning algorithm from a collection of annotated texts, where each text is usually represented by a vector of words (bag of words), skip-grams or n-grams, in combination with other types of semantic features that try to form the syntactic structure of sentences, negation, intensification, irony or subjectivity. The most popular are classifiers based on :

- SVM (Support Vector Machines),
- Naive Bayes and
- KNN (K-Nearest Neighbor)
- LSA (Latent Semantic Analysis) and
- Deep Learning.

Our solution:

Our Sentiment Analysis API is using semantic approaches which is based on advanced natural language in all aspects of syntax, morphology, pragmatics, and semantics. Firstly engine generates a semantic-syntactic tree of the text, and over this, lexicon terms are applied to spread their values of polarity all along the tree, properly combining the values depending on the morphological category of the word and the syntactic relations that they are affected by.

In addition to the overall polarity of the text, the engine returns the polarity for segments of the text or the word groups, in 6 possible levels: **negative** (N) and **positive** (P), **very negative** (N+) and **very positive** (P+), then comes **neutral** (NEU) and **none** (NONE) for an event where no polarity is involved.

Let take this example, this text is given:

We don't like the astonishing rise of the stock in this week.
Engine returns a total of **N** polarity, and also shows that this portion of sentence "*the astonishing rise of the stock this week*" has **P+** polarity. It also marks the phrase as **subjective**. This information shows a more detailed and accurate interpretation of the message that the text wants to convey.

If instead of "*I don't like*", the phrase was "*I don't like at all*", the overall value would be **N+** polarity (that's very negative), which is an important to note and analyze.

Aspects are the future:

Now, the upcoming trend is to move a step forward from the analysis of the total polarity of documents. The market demand is of a detailed fine-grained analysis of the messages that are expressed in the given text. So, the actual task evolves into **aspect-based sentiment analysis** (ABSA), whose objective is the classification and extraction of feelings and opinions on each specific aspect, which can be a topic label, a particular entity, a concept, or, in general, any dimensional analysis of an interest.

Our solutions provide the ability to **process high volumes of data** with **minimal delay**, **high accuracy**, **consistency** and **low cost**, which complements the human analysis in many scenarios.

1.2 PROBLEM STATEMENT

With the ever increasing volume of online reviews by the passing time, it is becoming very difficult for the customer to read all of them to form a perspective about the restaurant to be visited. Moreover, the customer may miss out some of the features of a particular restaurant like food, ambience, and service etc.

In particular this NLP application explicitly addresses the issue of not being able to frame an opinion about the restaurant from a large amount of restaurant reviews by analyzing the semantics of the reviews and summarizing large volumes of data into a small paragraph which will be as per the customer convenience.

1.3 OBJECTIVE

To perform sentiment analysis of the user reviews as a hybrid of lexicon based and machine learning algorithm. The aspects used for rating are: food, service, atmosphere, value. So, our focus will be on each aspect from the reviews provided by the customer. For each rated aspect, predictor models are to be trained independently. With the help of Data Cleaning and Text Extractor we will break review texts into a set of text fragments unigram words for analyses. Besides unigrams in the review text, we will also use the counts of positive and negative polarity words and their differences. Polarity labels will be obtained from a dictionary of about 700 words. We will use Machine learning to predict ordinal ratings from textual data and use MaxEnt

classification to minimize classification error. Then we partition the data into 80% for training and 10% for validation and 10% for testing. Random distribution ensures training and test data are identical.

1.4 METHODOLOGY

- **Web Crawler:** This step involves dynamic data extraction from online source, say a website. The data crawled by the web spiders will then be dumped into a csv file in our system.

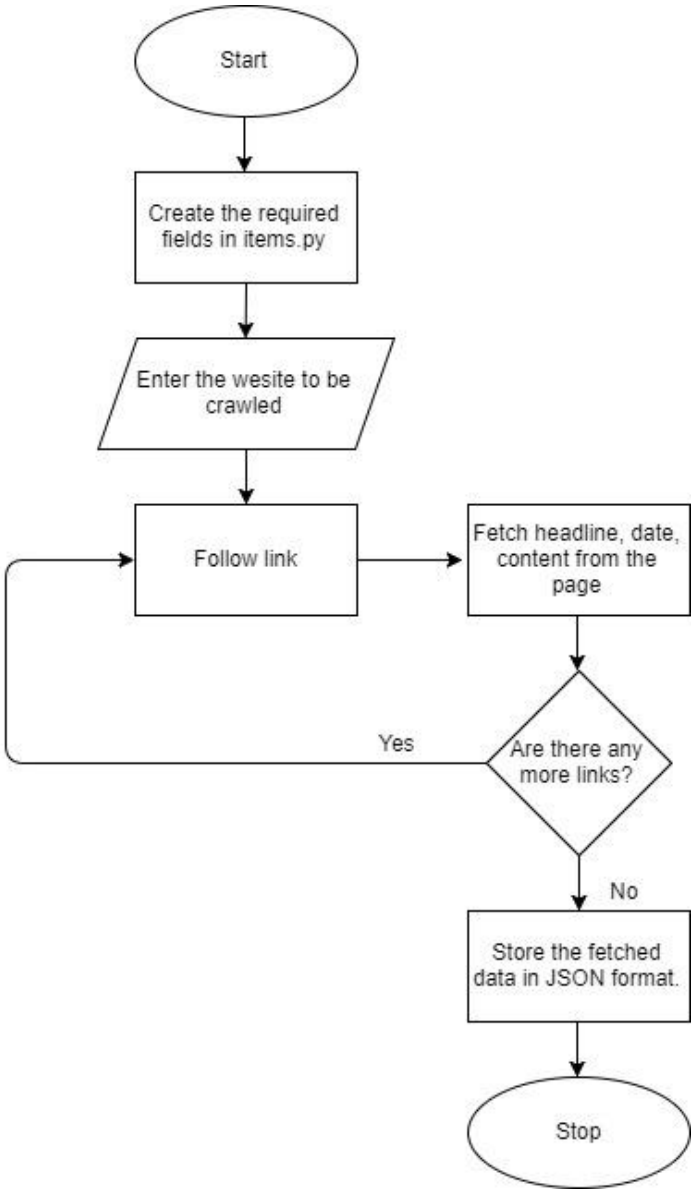


Figure 7 : Web Crawler

Web Crawler fetches each link on the page i.e. crawls through each restaurant on the page, and stores the links as a list of strings which will be traversed one by one.

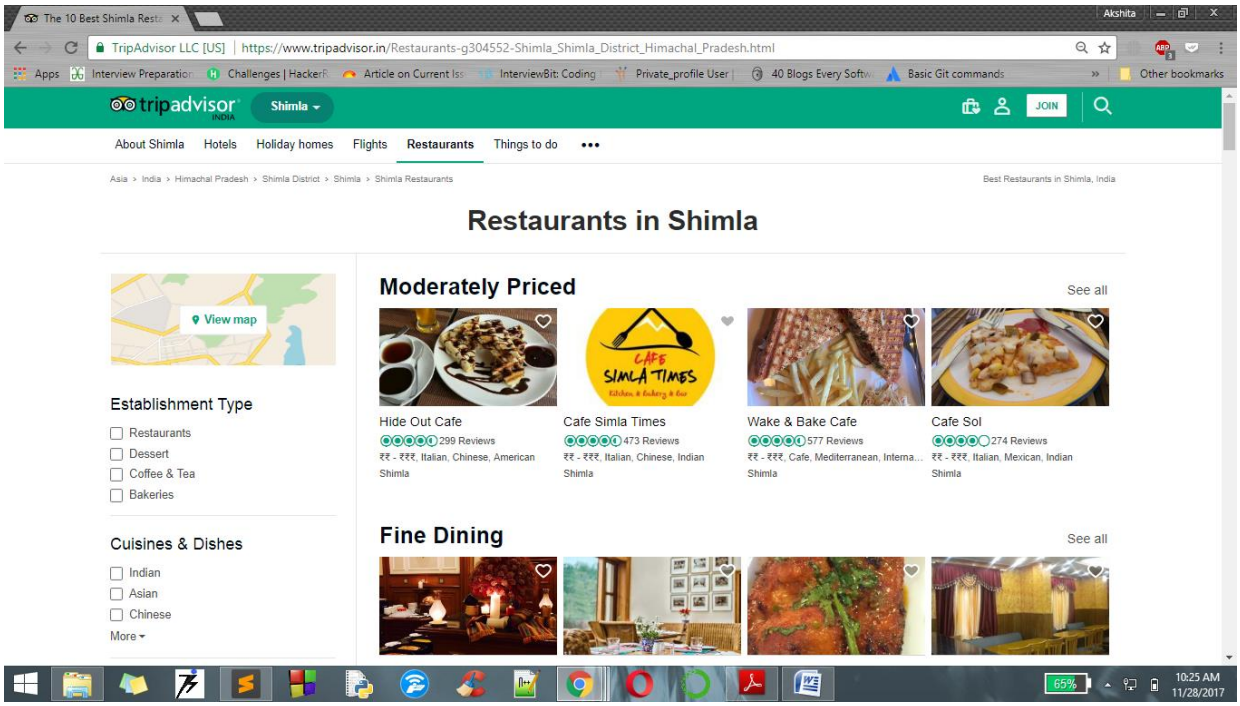


Figure 8 : TripAdvisor

Now, the list will be traversed and the crawler will move to each restaurant individually. The reviews will now be fetched. All the reviews given by the different users will be joined and fetched as a single string of reviews

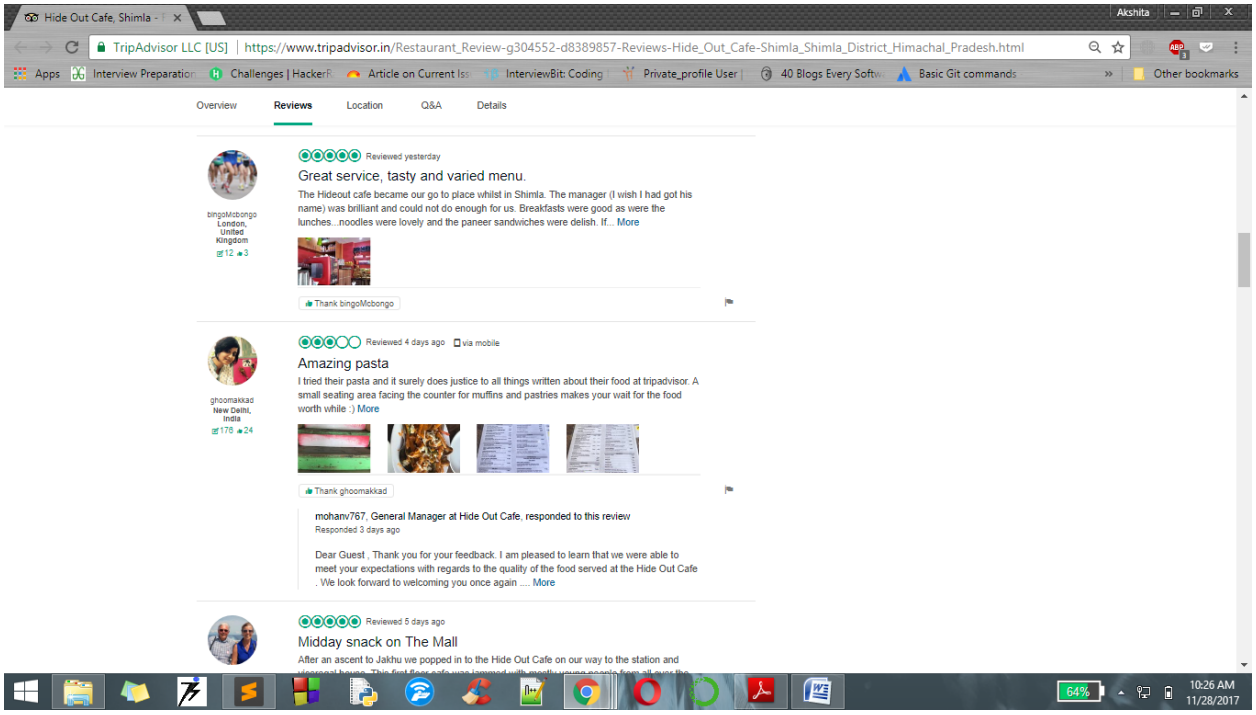
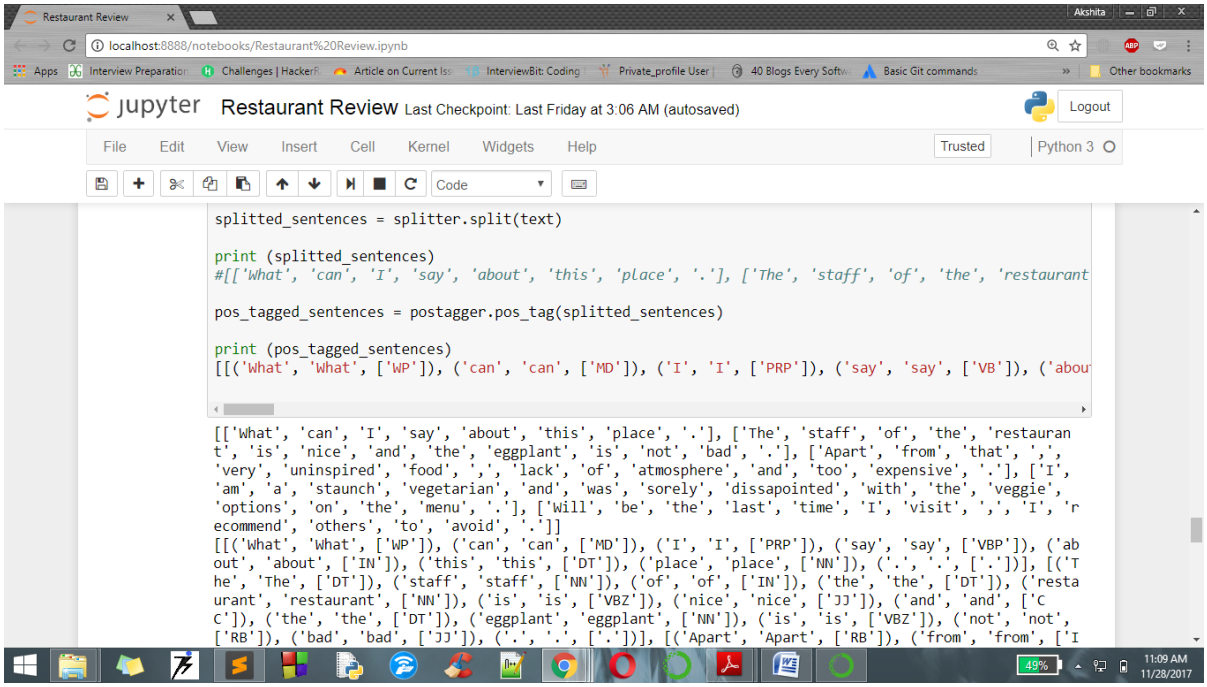


Figure 9 : Reviews of each restaurant

- **Data Cleaning:**

  Data cleaning is the process of altering/changing data in a given storage resource to make sure that it is correct and accurate. There are many ways to process data cleansing through various data storage architectures and software; most of them are centered on the precise review of data sets and also on the protocols associated with any particular data storage technology.



- **Sentiment Analyzer:** The first job after data dumping is to analyze the sentiment of a large number of text reviews corresponding to each restaurant. For this purpose we will train our data using SVM classifier.

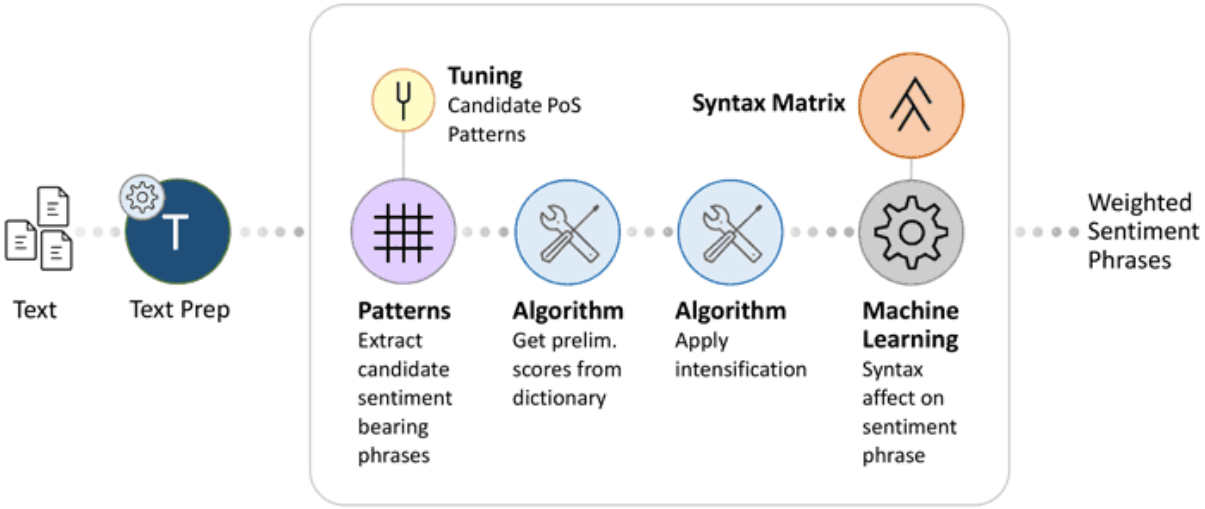


Figure 10 : Framework of Sentiment Analyzer [10]

- **Text Summarizer:** The next step would be to summarize the reviews of almost five-hundred lines into a paragraph of 5-6 lines which is more readable. We are going to use

different text summarizer for this purpose and at the end we will compare all those to find out the best summarizer.

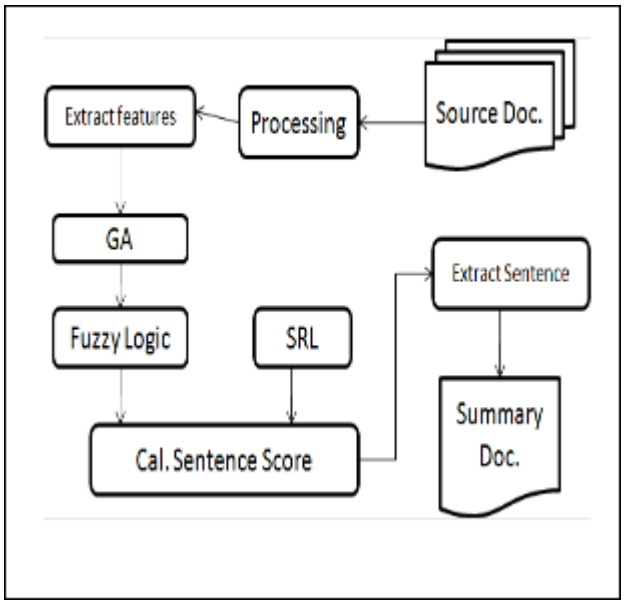


Figure 11: : Working of a Text Summarizer [11]

- **Front-end application:** The front-end application will be designed at the end using flask (a python framework).

CHAPTER 2: LITERATURE SURVEY

2.1 Title: "Building a Sentiment Summarizer for Local Service Reviews"

It uses data mining, content analysis and indexing for summarization of opinions and display them in easy to process manner for Information Application Systems. It shows classification on the basis of raw-score, purity and user generated rating. Later they have compared four systems review-label, raw-score, max-ent, and max-ent-review-label. Finally they have combined the results from dynamic and static aspect extractor. They have done sentiment analysis on various projects like Department Store, Children's Barber Shop, Greek Restaurant, and Hotel and Casinos.

2.1.1. Proposed System Architecture

Proposed system and its architecture contains following modules:

1. Text Extractor
2. Sentiment Classifier
3. Dynamic and Static Aspect Extractor
4. Aggregator and Summarizer

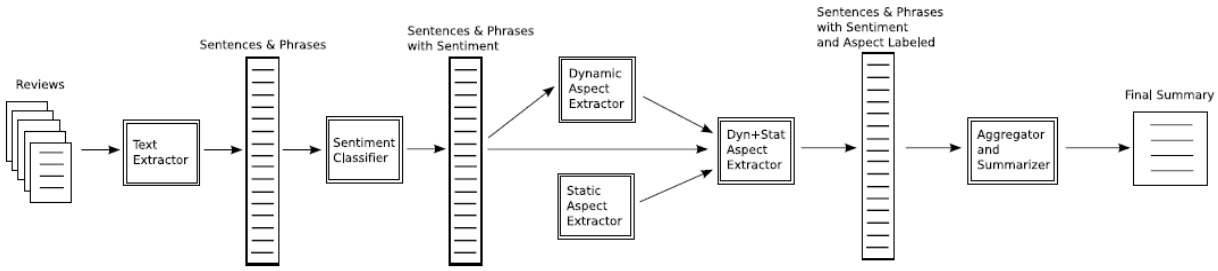


Figure 12 : System Overview

2.1.2. Conclusion:

In this paper they have presented architecture for summarizing the sentiments. The resulting system is highly accurate for most often queried services at the same time it is also sufficiently general to produce quality summaries for all types of services. The main technical contributions in this paper are the new sentiment models that leverage context and user-provided labels for improving sentence level classification as well as a hybrid aspect extractor and summarizer that combines supervised and unsupervised methods to improve system's accuracy. In the future we plan to modify the system for different products, that's a domain which has been well studied in the past.

Like in services section, they believe that hybrid models can improve system's performance because it has been seen that generally there exists a pattern that very few products account for maximum number of reviewed queries (e.g., electronic items). After precise observation, it has been noted that there is a set of aspects that is common across most products, such as customer service, warranty, and value, which can definitely be utilized to improve the performance of the less queried products.

We've also planned to run user interface studies to find the best mode of delivery of aspect-based sentiment summarization for both desktop and mobile computing platforms. They have observed that by changing the number of aspects shown and also the granularity of the associated text, there can be large changes can be seen in the summary. More investigation of semi-supervised and active learning methods for aspect classification may provide a mechanism for further reducing the amount of labeled data required to produce highly accurate and coarse-grained aspects.

2.2 Title: "Beyond the Stars: Improving Rating Predictions using Review Text Content"

Since keyword searches typically do not provide good results, as the same keywords are routinely appearing in good and as well as in bad reviews. Yet another challenge in understanding reviews is that a reviewer's total rating might be largely reflective of product features in which the search user is not at all interested. But, identifying structured information from free-form text is a challenging task as users routinely enter informal text with poor spelling and grammatical errors. Mostly it happens that reviews contain anecdotal text and similar information, which is not very useful or usable information for automatic processing.

Their work takes the approach of combining machine learning, natural language processing and collaborative filtering to harness the wealth of detailed information available in the web reviews.

Their approach is of combining natural language processing with machine learning and collaborative filtering to analyze detailed information available in web reviews. They have captured reviews in the following four categories that are positive, negative, neutral, and conflict category, where user is comparing and contrasting good and bad aspects.

They have used support vector machines (SVMs) for manual annotated data. Rating has been done on service, food, ambience, price, anecdotes and miscellaneous. On comparison they have seen that textual information gives much better results in form of general and personalized review score's predictions rather than the ones derived from the star rating given in numerical form by the users.

They have used Pearson Correlation Coefficient and Mean Squared Error accuracy metric to evaluate prediction technique.

2.2.1. Conclusion:

In this paper the main contribution is the analysis of the impact or effect of text-derived information in anticipating the rating of a review in a re-commendation system. We show that both topic and sentiment information at the sentence level are useful information to leverage in a review. To the best of our knowledge, this is the first time that the textual component of the review has been considered in such systems, and that user reviews are analyzed and classified at the sentence level. We are investigating additional refinements to our text-based recommendation, including better text classification strategies, allowing
users to get recommendations on specific aspect of restaurants such as food or ambience, and soft clustering-based approaches that group users based on their reviewing styles and interest similarities. In addition, we are interested in the impact of text classification on search over reviews and are implementing tools that allow users to search reviews using category and sentiment information.

2.3 Title : "Capturing the stars: predicting ratings for service and product reviews"

Their work addresses the task of automatically predicting ratings, for given aspects of a textual review, by assigning a numerical score to each evaluated aspect in the reviews. They have handled this task as both a regression and a classification modeling problem and explore several combinations of syntactic and semantic features. Results suggest that classification techniques perform better than ranking modeling when handling evaluative text.

They have considered rating of five aspects food, atmosphere, value, service and overall experience. They have used supervised learning methods to train predictive models and improve their performance by optimizing classification predictions.

They have used numerical regression technique viz. neural network, an ordinal regression technique viz. PRank algorithm, and classification technique viz. MaxEnt classifiers. They have calculated rank loss as the difference between actual and predicted rating. All the results quoted in this paper are averages of 10–fold cross– validation over 6,823 review examples. They have treated aspect rating predictors as independent of each other. They process only nouns, verbs, and adjectives.

There are several possible approaches to such a regression problem:

1. The most obvious approach is numeric regression. It is implemented with a neural network trained using the back–propagation algorithm.

2. Ordinal regression can also be implemented with multiple thresholds (r − 1 thresholds are used to split r ranks). This is implemented with a Perceptron based ranking model called Prank.

3. Since rating aspects with values 1, 2, 3, 4 and5 is an ordinal regression problem it can also be interpreted as a classification problem, with one class per possible rank. In this interpretation, ordering information is not directly used to help classification. Our implementation uses binary one-vs-all Maximum Entropy (MaxEnt) classifiers.

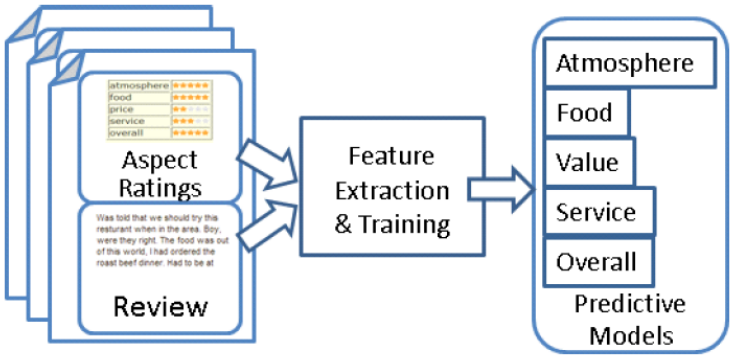


Figure 13: Predictive model training

Aspects	Ratings	Reviews
Atmosphere Food Value Service Overall	★★☆☆☆ ★★☆☆☆ ★★★☆☆ ★★☆☆☆ ★★☆☆☆	Heavy, uninspired food, eaten under appall of cigarette smoke. Very slow service, though not unfriendly. There are many better restaurants in Ashland. Not recommended.
Atmosphere Food Value Service Overall	★★★☆☆ ★★★☆☆ ★★★☆☆ ★★★★☆ ★★★☆☆	I'll have to disagree with Ms. Kitago's take on at least one part of the evening. I believe the chicken Tikka Marsala was slightly dry. Decent portion, but not succulent as i am accustomed to. In addition, the Gulub Jaman is served cold, anathema to this diner. I will agree with Ms. K that the mango lassi was delicious, but overall I believe her review was slightly inflated

Figure 14: Example of ratings with partial support in the textual format

2.3.1. Conclusion:

Textual reviews for different products and services are abundant. Still, when trying to make a buy decision, getting sufficient and reliable information can be a daunting task. In this work, instead of a single overall rating we focus on providing ratings for multiple aspects of the product/service. Such ratings must be deduced from predictive models because most textual reviews are rarely accompanied by multiple aspect ratings. Several authors in the past have studied this problem using both classification and regression models. In this work we show that even though the aspect rating problem seems like a regression problem, maximum entropy classification models perform the best. Results also show a strong inter–dependence in the way users rate different aspects.

2.4 Title : "Automatic Identification of Pro and Con Reasons in Online Reviews "

Their focus is on extracting the reasons of the opinions, which may themselves be in the form of either fact or opinion. They have studies opinion at three different levels word level, sentence level and document level. Pros in a product review are sentences that describe reasons why an author of the review likes the product. Cons are reasons why the author doesn't like the product. They have successfully applied Maximum Entropy classification in many tasks of natural language processing, like Semantic Role labeling, Question Answering, and Information Extraction. Their classification uses three types of features: lexical features, positional features, and opinion bearing word features. They have divided data: 80% for training, 10% for development, and 10% for test for experiments.

Class symbol	Description
PR	Sentences related to pros in a review
CR	Sentences related to cons in a review
NR	Sentences related to neither PR nor CR

Figure 15: Classes defined for the classification

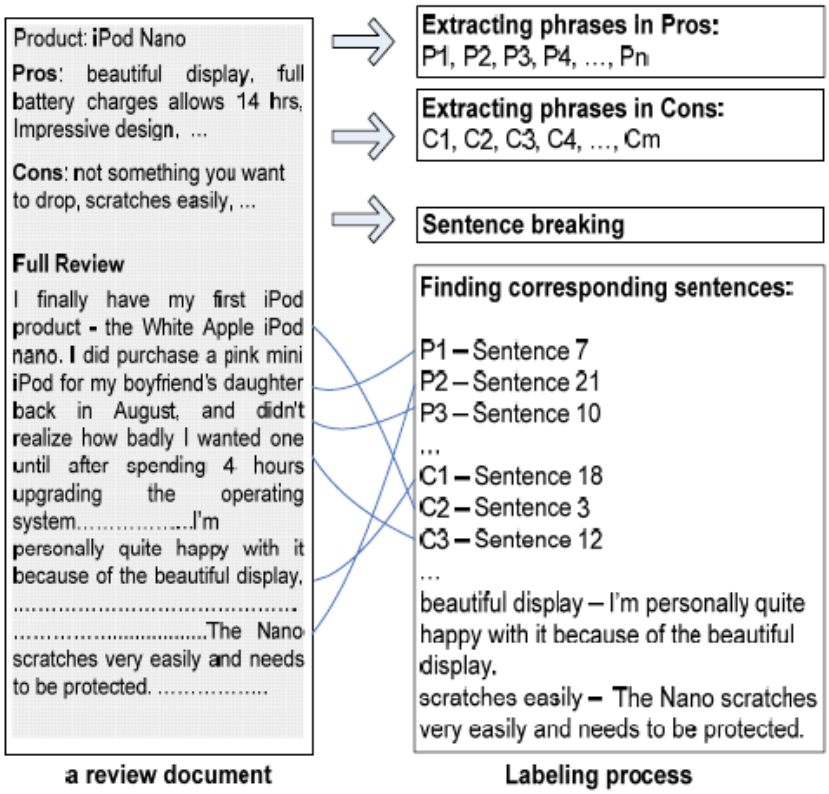


Figure 16 : The automatic labeling process of pros and cons sentences in a review.

2.4.1. Conclusion:

This paper proposes a framework for identifying one of the critical elements of online product reviews to answer the question, "What are reasons that the author of a review likes or dislikes the product?" They believe that pro and con sentences in reviews can be answers for this question. They've present a technique that automatically labels a huge set of pros and cons sentences in online reviews, by using clue phrases for pros and cons inepinions.com in order to train their system. They've applied it to label sentences both on epinions.com and complaints.com.

To investigate there liability of their system, they've tested it on two extremely different review domains, mp3 player reviews and restaurant reviews. Their system with the best feature selection performs 71% F-score in the reason identification task and 61% F-score in the reason classification task. The experimental results further show that pro and con sentences are a mixture of opinions and facts, making identifying them in online reviews a distinct problem from opinion sentence identification.

Finally, they've also applied the resulting system to another review data in complaints.com in order to analyze reasons of consumers' complaints. In the future, they plan to extend their pros and cons identification system on other sorts of opinion texts, such as debates about political and social agenda that they can find on blogs or newsgroup discussions, to analyze why people support a specific agenda and why people are against it.

2.5 Title : "Sentiment Analysis: Capturing Favorability Using Natural Language Processing"

It shows a sentiment analysis approach to extract sentiments associated with polarities of positive or negative for specific subjects from a document, instead of classifying the whole document into positive or negative. They have done sentiment analysis, favorability analysis, text mining, and information extraction.

They've assumed that if they can detect favorable and unfavorable opinions they can get a powerful functionality for analysis of competition and marketing and they believe they can also detect unfavorable rumors related to risk management. Analysis of statements expressing sentiments is more reliable compared to the overall opinion.

Therefore, instead of analyzing the favorability of whole context they try to extract each statement on its particular favorability. Their sentiment analysis involves identification of: Polarity and strength of the expressions , Sentiment expressions, and their relationship to the subject under consideration. They find sentiment expression of the given subject and determine the polarity of the sentiments. To identify sentiment expressions and analyze their semantic relationships with the subject term, natural language processing is used. Some cases failed due to complex sentence structure, in the input context that negates the local sentiment for the whole, and they are not due to failures of our syntactic parser. So they restrict the output of ambiguous cases that tend to be negated by predicates at higher levels,

Like interrogative sentences, if-else clause, and sentiments in noun phrase. They obtain 99% precision in data set by eliminating ambiguous sentiments.

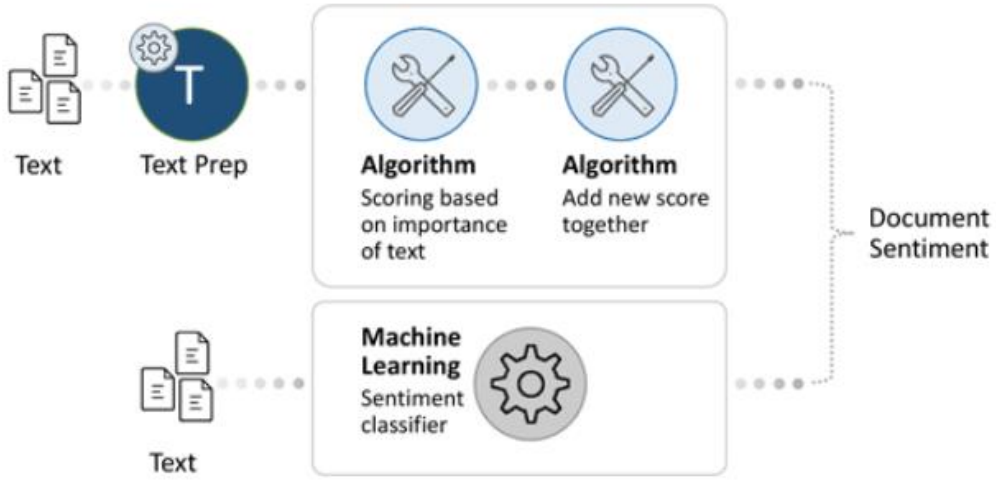


Figure 17: System design [10]

2.5.1 Conclusion

The initial experiments resulted in about 95% precision and roughly 20% recall. However, as they've expanded the domains and data types, they've observed some difficult data for which the precision can go down to about 75%. Point to be noted, such data usually contains well-written texts like news articles and descriptions in some official organizational Web pages. Since those texts often contain long and complex sentences, our simple framework finds them difficult to deal with. As seen in the examples, most of the failures are due to the complex structures of sentences in the input context that negates the local sentiment for the whole, and they are not due to failures of our syntactic parser. For example, a complex sentence such as:

"It's not that it's a bad camera" confuses our method. It is noteworthy that failures in parsing sentences do not damage the precision in our approach. Also, it allows them to classify the ambiguous cases, it is done by identifying features in sentences, like inclusion of interrogatives and if-clauses. Thus precision can be maximized by totally removing ambiguous cases like these, mostly for applications that prefer precision rather than recall.

2.6 Title: "Sentiment analysis using support vector machines with diverse information sources"

This is an approach to sentiment analysis which uses (SVMs) Support Vector Machines that helps in bringing together diverse sources that contain potentially pertinent information. Detecting the tone of information, it can be positive or negative, this is an important step. Two word phrases conforming to particular part-of-speech templates representing possible descriptive combinations are used. They have used POI (pointwise mutual information) and SO (semantic orientation). They have used Osgood's semantic orientation with WordNet to evaluate three values that correspond to the activity (active or passive), potency (strong or weak), and the evaluative (good or bad).

Texts were annotated by hand using the Open Ontology Forge annotation tool. It includes value-phrase preposition topic entity as a distinct class—often. The accuracy value represents the percentage of test texts which were classified correctly by the model. The best results are obtained using such hybrid SVMs, which yield scores of 84.6% accuracy on the 3-fold experiments and 86.0% accuracy on the 10-fold experiments. The simple lemmas model obtains an average score of 84% and the simple unigrams model obtains 79.75%.

Figure 18 : n-grams

Conclusion :

The method introduced in this paper allows many different methods of assigning semantic values to words and phrases within a text to be used in a more useful way than was previously possible, by incorporating them as features for SVM modeling, and for explicit topic information to be utilized, when available, by features incorporating such values. Combinations of SVMs using these features in conjunction with SVMs based on unigrams and lemmatized unigrams are shown to outperform models which do not use these information sources. The approach presented here is flexible and suggests promising avenues of further investigation.

2.7 Title : "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text"

VADER (Accuracy = 0.96) performs much better than individual human raters (Accuracy = 0.84) by correctly classifying tweets sentiment's into positive, neutral, or negative classes.

Sentiment lexicons have been used like a list of lexical features (e.g., words) that are labeled according to their semantic orientation as either positive or negative and (SVMs) Support Vector Machine algorithms, Naive Bayes and Maximum Entropy. They have tried Word-sense disambiguation which indicates to the process of identifying which sense of a word is used in a sentence since words have multiple meanings, therefore improving sentiment analysis performance by understanding deeper lexical properties. They have compared the VADER sentiment lexicon to other seven quite-established sentiment analysis lexicons:

- Linguistic Inquiry Word Count (LIWC),
- General Inquirer (GI),
- Affective Norms for English Words (ANEW),
- SentiWordNet (SWN),
- SenticNet (SCN),
- Word-Sense Disambiguation (WSD) using WordNet, and
- the Hu-Liu04 opinion lexicon.

VADER has many advantages; it is quick and economic computationally without taking a toll on accuracy.

2.7.1 Conclusion

VADER is (Valence Aware Dictionary for sEntiment Reasoning). They've reported the systematic development and evaluation of VADER. Combining of qualitative and quantitative methods, they've constructed and empirically validated *gold standard* list of lexical features (along with their associated sentiment intensity measures) which are specifically tuned to sentiment in microblog-like contents.

They've then combined all these lexical features by giving considerations to 5 very general rules that embody syntactical and grammatical conventions used for emphasizing and expressing of sentiment intensity. The results they got are encouraging as well as quite remarkable. VADER performed equivalent to (and in most cases, *better than*) eleven other highly regarded sentiment analysis tools. Its results highlight the developments that can be made in computer science when the human is believed as a central part of the development process.

CHAPTER 3: SYSTEM DEVELOPMENT

3.1) SOFTWARE REQUIREMENTS:

- Python: 3.6.1
- Anaconda 4.4.0
- NLTK
- Scikit
- Scrapy

3.2) HARDWARE REQUIREMENTS:

- **System Requirements:**
  - CPU: 2.2 GHz Processor and above
  - RAM: 4 GB or above
  - OS: Windows 7 or above

3.3) PROPOSED MODEL

3.3.1. TECHNOLOGIES USED

- **Python**

  Python is a dynamic, object-oriented, and interpreted (that's bytecode-compiled) language. There is no data-type for declarations of variables, parameters, functions, or methods in the source code. It makes the source-code precise and flexible, and you lose the compile-time type checking of the source code. Python has the ability to tracks the data-type of all values during runtime and python flags the code that doesn't make sense when it runs.

- **Anaconda**

  Anaconda is a Python distribution that is precisely popular for data analysis and scientific computing work. Open source project developed by Continuum Analytics Inc. available for Mac OS X, Windows and Linux including many popular packages like: SciPy, NumPy, Pandas, Matplotlib, IPython, and Cython which includes Spyder. A Python development environment includes conda, which is a platform-independent package manager.

- **Scrapy**

  Scrapy is a free and open source web crawling framework in Python. It is a framework that provides different classes for performing different tasks, therefore eases the functioning and refines the code at the same time. Spiders are self-contained crawlers which have been given a set of instructions and they work accordingly. Scrapy's project architecture is built around spiders. It can also be used to extract data using APIs and also as a general purpose webcrawler

- **NLTK(Natural Language Toolkit)**

  NLTK is a major platform for building Python programs that work with human language data. It provides easy-to-use interfaces and an excellent suite for various processes like text processing libraries for tokenization, classification, tagging, stemming, parsing, and also semantic reasoning. It helps as wrappers for industrial strength NLP (Natural Language Processing) libraries, and an active discussion forum. Natural Language Processing is applied on text messages, emails, social media posts, search queries, legal and medical records, and many more areas to obtain efficient results.

- **Scikit**

  Scikit-learn (formerly known as **scikits.learn**). It is a free software machine learning library for Python. It implements many regression, classification, and clustering algorithms also including support vector machines(SVMs), and random forests.It is also designed to inter-operate with the Python scientific and numerical libraries like SciPy and NumPy.

3.3.2. OTHER LIBRARIES:

These are some of the basic libraries that transform Python from a general purpose object-oriented programming language into a powerful and a robust tool for multi-purpose activities like data visualization and data analysis. Also called the SciPy Stack, they're the foundation with help of which more specialized tools are built.

- **NumPy** is a foundation library used for scientific computing in Python, and there are many other libraries which use NumPy arrays as basic inputs and outputs. It generally introduces the objects for matrices and multi-dimensional arrays, it also works with routines that allow developers in python to operate on advanced statistical and mathematical functions on those arrays even with very little code.

  SciPy is built on NumPy by adding a collection of algorithms and high-level commands for visualizing and manipulating data. SciPy includes functions for numerical computation for integrals, optimization, solving differential equations and much more.

- **Pandas** adds data structures and other tools which are designed for real-time data analysis in statistics, finance, engineering, and social sciences. Pandas works quite well with messy, incomplete, and unlabeled data (that's the data you generally encounter in the real world), and provides tools for merging, shaping, slicing datasets, and reshaping.

- **IPython**, it is a framework that extends the functionality of Python's interpreter which is interactive with an inbuilt shell that adds rich media, introspection, tab completion, shell syntax and also command history retrieval. It acts as an embedded interpreter for your programs that is really useful for debugging. If you have ever used MATLAB or Mathematica, you will feel quite comfortable with IPython.

  Matplotlib is the standard Python library, used for creating 2D plots and graphs. It's has low-level features, that meaning it requires many commands to generate nice-looking figures and graphs in comparison to with some more advanced libraries. It's flip side is flexibility. That is with enough commands, user can make just about any kind of graph one wants with matplotlib.

3.3.3. SYSTEM DESIGN:

The first step would be to fetch the data using a Web-Crawler and dump it into an excel file (CSV file) or store it in JSON format. The data fetched now undergoes various data

cleaning operations using the NLTK. Hence, the review data is cleaned. Then to check for the waitage of all the important and lesser important words, we create a document term matrix, hence removing the least important as they are considered of least important in the matrix. The next step would be split the data into two different sets of test and train set. The data is is split such that the training data is almost 80% or close and the rest will be categorized as the test data ,which will be used to test my trained data for the labels.

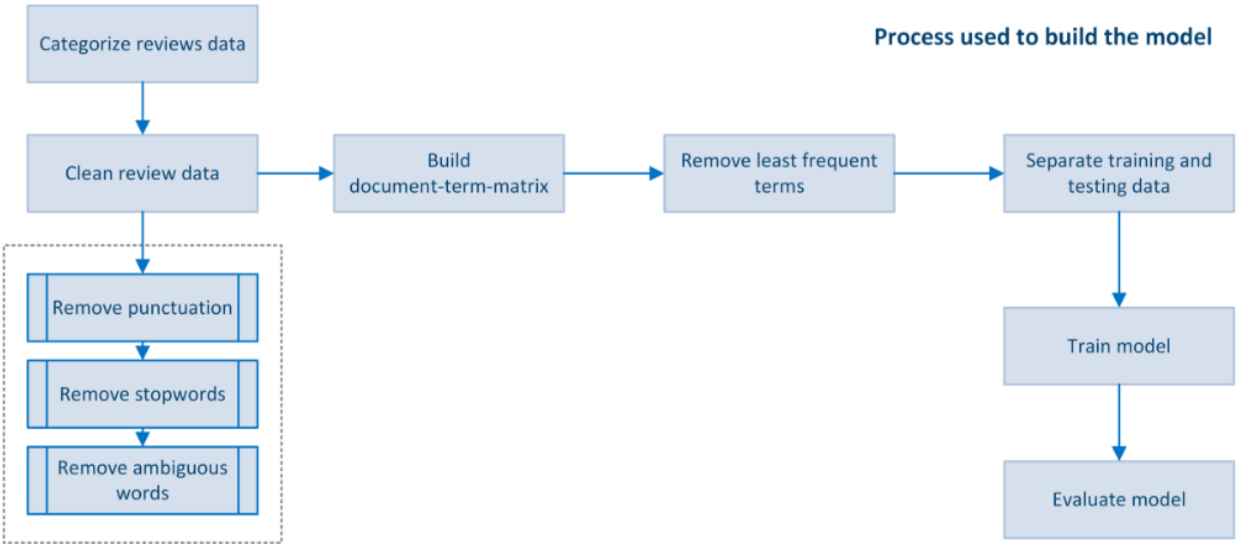


Figure 19 : Proposed System Model

Use case diagram

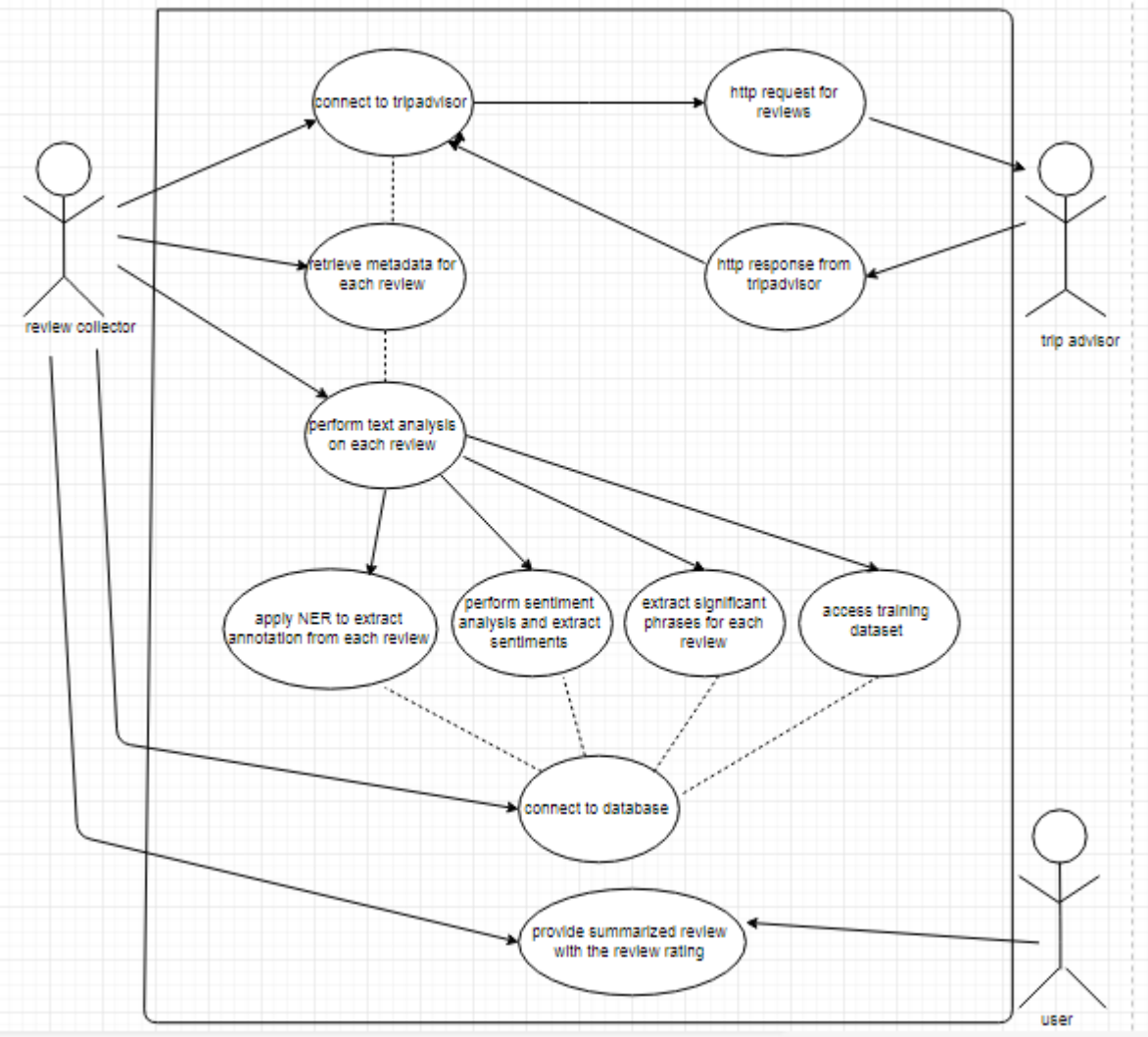
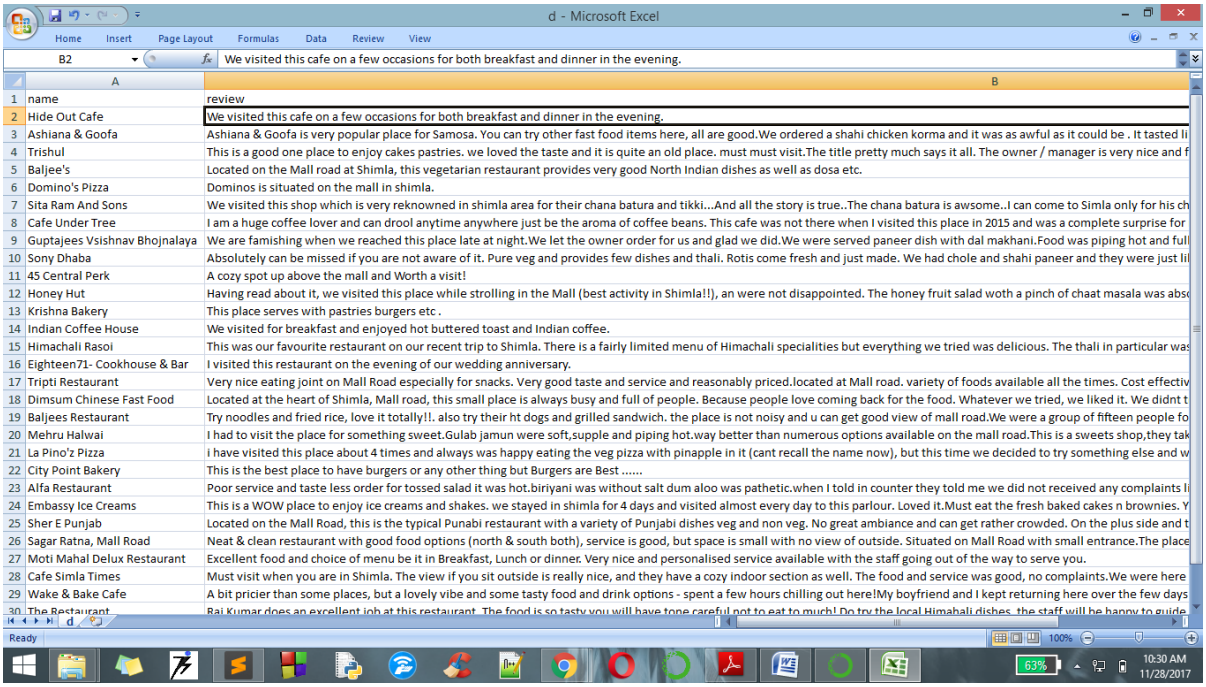


Figure 20: Use case diagram

3.3.4. DATA SET:

We extracted our corpus of over 100 restaurant reviews from TripAdvisor Shimla. This corpus contains around 100 restaurants, with associated structured information about location, cuisine type and a set of reviews. Reviews contain structured metadata about star rating, and date along with text. Generally reviews are small; an average user's review has 5.28 sentences. The reviews are written by 50-60 distinct users, for whom we only have unique username information. The data set is sparse: that's because restaurants generally have only a few reviews, with 100 restaurants having more than 10 reviews; and users mostly review few restaurants, with only 299 (non-editorial) users who would have reviewed more than 10 restaurants.



	A	B
1	name	review
2	Hide Out Cafe	We visited this cafe on a few occasions for both breakfast and dinner in the evening.
3	Ashiana & Goofa	Ashiana & Goofa is very popular place for Samosa. You can try other fast food items here, all are good.We ordered a shahi chicken korma and it was as awful as it could be . It tasted li
4	Trishul	This is a good one place to enjoy cakes pastries. we loved the taste and it is quite an old place. must must visit.The title pretty much says it all. The owner / manager is very nice and f
5	Baljee's	Located on the Mall road at Shimla, this vegetarian restaurant provides very good North Indian dishes as well as dosa etc.
6	Domino's Pizza	Dominos is situated on the mall in shimla.
7	Sita Ram And Sons	We visited this shop which is very reknowned in shimla area for their chana batura and tikki...And all the story is true..The chana batura is awsome..I can come to Simla only for his ch
8	Cafe Under Tree	I am a huge coffee lover and can drool anytime anywhere just be the aroma of coffee beans. This cafe was not there when I visited this place in 2015 and was a complete surprise for
9	Guptajees Vsishnav Bhojnalaya	We are famishing when we reached this place late at night.We let the owner order for us and glad we did.We were served paneer dish with dal makhani.Food was piping hot and full
10	Sony Dhaba	Absolutely can be missed if you are not aware of it. Pure veg and provides few dishes and thali. Rotis come fresh and just made. We had chole and shahi paneer and they were just lil
11	45 Central Perk	A cozy spot up above the mall and Worth a visit!
12	Honey Hut	Having read about it, we visited this place while strolling in the Mall (best activity in Shimla!!), an were not disappointed. The honey fruit salad woth a pinch of chaat masala was abs
13	Krishna Bakery	This place serves with pastries burgers etc .
14	Indian Coffee House	We visited for breakfast and enjoyed hot buttered toast and Indian coffee.
15	Himachali Rasoi	This was our favourite restaurant on our recent trip to Shimla. There is a fairly limited menu of Himachali specialities but everything we tried was delicious. The thali in particular was
16	Eighteen71- Cookhouse & Bar	I visited this restaurant on the evening of our wedding anniversary.
17	Tripti Restaurant	Very nice eating joint on Mall Road especially for snacks. Very good taste and service and reasonably priced.located at Mall road. variety of foods available all the times. Cost effectiv
18	Dimsum Chinese Fast Food	Located at the heart of Shimla, Mall road, this small place is always busy and full of people. Because people love coming back for the food. Whatever we tried, we liked it. We didnt t
19	Baljees Restaurant	Try noodles and fried rice, love it totally!!. also try their ht dogs and grilled sandwich. the place is not noisy and u can get good view of mall road.We were a group of fifteen people fo
20	Mehru Halwai	I had to visit the place for something sweet.Gulab jamun were soft,supple and piping hot.way better than numerous options available on the mall road.This is a sweets shop,they tak
21	La Pino'z Pizza	i have visited this place about 4 times and always was happy eating the veg pizza with pinapple in it (cant recall the name now), but this time we decided to try something else and w
22	City Point Bakery	This is the best place to have burgers or any other thing but Burgers are Best .....
23	Alfa Restaurant	Poor service and taste less order for tossed salad it was hot.biriyani was without salt dum aloo was pathetic.when I told in counter they told me we did not received any complaints li
24	Embassy Ice Creams	This is a WOW place to enjoy ice creams and shakes. we stayed in shimla for 4 days and visited almost every day to this parlour. Loved it.Must eat the fresh baked cakes n brownies. Y
25	Sher E Punjab	Located on the Mall Road, this is the typical Punabi restaurant with a variety of Punjabi dishes veg and non veg. No great ambiance and can get rather crowded. On the plus side and t
26	Sagar Ratna, Mall Road	Neat & clean restaurant with good food options (north & south both), service is good, but space is small with no view of outside. Situated on Mall Road with small entrance.The place
27	Moti Mahal Delux Restaurant	Excellent food and choice of menu be it in Breakfast, Lunch or dinner. Very nice and personalised service available with the staff going out of the way to serve you.
28	Cafe Simla Times	Must visit when you are in Shimla. The view if you sit outside is really nice, and they have a cozy indoor section as well. The food and service was good, no complaints.We were here
29	Wake & Bake Cafe	A bit pricier than some places, but a lovely vibe and some tasty food and drink options - spent a few hours chilling out here!My boyfriend and I kept returning here over the few days
30	The Restaurant	Raj Kumar does an excellent job at this restaurant. The food is so tasty you will have tone careful not to eat to much! Do try the local Himahali dishes, the staff will be happy to guide

Figure 21 : Data fetched and dumped in a csv file

PARAMETERS FETCHED:

- Name of Restaurant : The name of the restaurant is the first thing to be fetched. The names are stored in a particular column in the csv file.
- Reviews : All the customer reviews are fetched as a single string under the heading review.
- Rating : The rating fetched plays an important role in depicting the sentiment of the reviews.
- Location: The location is a selecting factor whether or not people will visit the place despite of the ambience and the food quality.

CHAPTER 4: ALGORITHM

Step 1: Fetch data from TripAdvisor website using a WebCrawler.

Step 2: Store that data into excel files. For text summarizations go to step 9. For sentiment analyses go to step 3.

Step3: Perform data-cleaning by removal of unwanted words.

Step 4: Sentiment Analysis is done by Natural Language Processing using the TF-IDF formula.

Step 5: TF-IDF rating corresponding every keyword is sent to SVM.

Step 6: By applying Machine-learning, use Support Vector Machine (SVM) technique data is classified into positive, negative and neutral keywords.

Step 7: Comparing review rating of different restaurants.

Step 8: Final sentiment analysis output is sent to user.

Step 9: After step 2, set of reviews per restaurant are used by TextRank algorithm for summarization of all the reviews for each restaurant.

Step 10: Final output is the summarized review per restaurant which is sent to the user.

CHAPTER 5: TEST PLAN

Testing is the process of evaluation of system and its components with an intention of finding whether it satisfies particular requirements or not. There is always a difference between actual and expected results. Therefore testing is necessary. Only by system execution we can identify gaps and errors and loopholes in the requirements.

To make sure system is error free, different level of system testing is applied:

1. Unit testing- Isolate each part and check if they are working correct individually. First test the high level modules then the lower level modules.
2. Integration testing- After testing the lowest level modules, we step up a level, and test the modules that were linked to previously examined modules. This is how bottom-up testing is done, followed by top-down testing.
3. System testing- When all components have been integrated, whole application is rigorously tested.
4. Acceptance testing- It is done to check whether application satisfies client's requirements or not and whether the application fulfills the required standards.
    a. Alpha testing- Unit testing and integration testing are together known as alpha testing.

Testing methods

1. White box- Internal logic and structure of code is tested in detail.
2. Black box- Only inputs are sent and outputs are analyzed by the tester without knowing system architecture and source code.

Validations- It is done when all levels of testing have been performed and all methods of testing have been implemented.

Limitations- SVMs take considerable time to execute.

Data that I have fetched from TripAdvisor website is of thirty restaurant reviews. It has been trained on a dataset of size of five hundred positive reviews and five hundred negative reviews.

I have divided the provided dataset into 10 percent for training and 90 percent for testing.

Further I’ve planned to test the developed model on data collected from different websites like Swiggy, Zomato and Foodpanda.

CHAPTER 6: RESULT AND PERFORMANCE ANALYSIS

While building a sentiment analysis system, we need to first split your data set into 3 sets:

1. **Training Set** (around 70%) - This set will be used for training your classifier.
2. **Validation Set** (around 15%) - This set would be used as a testing set while building and tuning the classifier.
3. **Test Set** (around 15%) - User would not use this test set until you have finalized your model and is ready for production. You would then run your system on this set to check the performance of your system i.e. whether your model generalized properly or not.

While building the system we would use the following measures:

1. **Cross-Validation Accuracy** - This accuracy would be calculated when performing k-fold cross-validation to tune the parameters.
2. **Accuracy** [Num. of Correct Queries / Total Num. of Queries] - User will use this to check the total accuracy of the system.
3. **F1-Measure** [2 * (Precision * Recall) / (Precision + Recall)] - Itis one of the most important measures that will tell user how the system is performing. Where Precision = True Positive / (True Positive + False Positive), and Recall = True Positive / (True Positive + False Negative
4. **Confusion Matrix** (Also called Error Matrix.) - This would also tell you how your system is performing and is similar to F1-Measure

Below are the graphs showing individual aspect analysis for different restaurants, restaurants on x axis and count on y axis. Individual aspects being rating, food, value and service. There is not a very high difference between these ratings for restaurant 1 and 5. Restaurant 3 and 4 shows

largest variations. Restaurant 2 shows medium variations. A lot of analysis still remains.

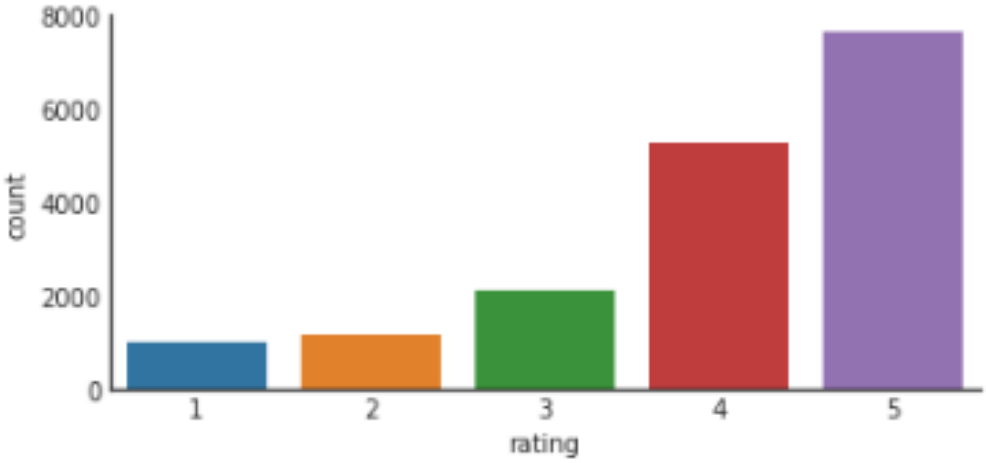


Figure 22: Rating for 1st restaurant

- This graph shows the rating of restaurants vs the count. This rating has been done on the basis of overall review given by the customers.

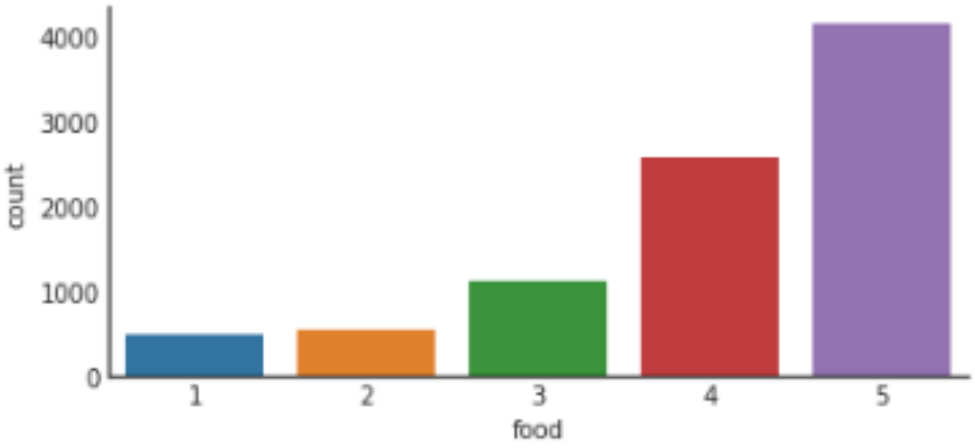


Figure 23: Rating for 2nd restaurant

- This is a graph of food vs count. The food rating for each restaurant has been calculated by analyzing the reviews for each and then finding out the sentiment associated with each along with the value provided to each sentiment.

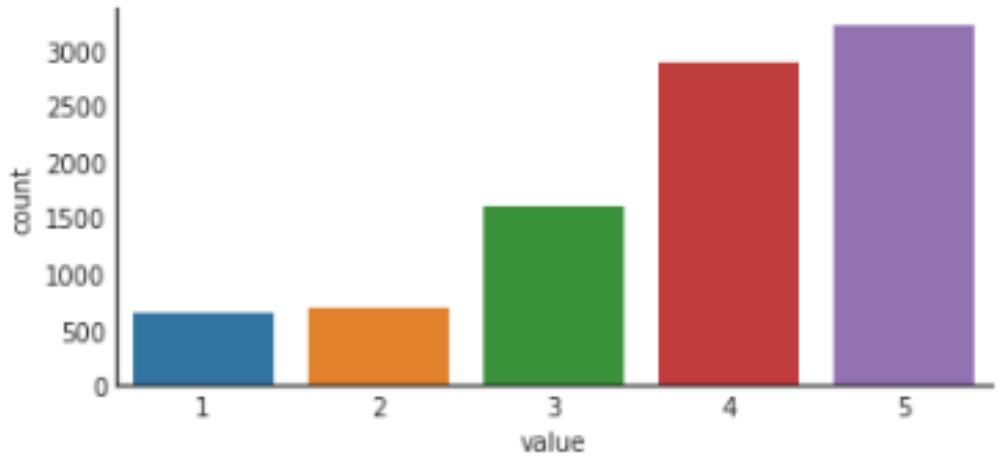


Figure 24: Rating for 3rd restaurant

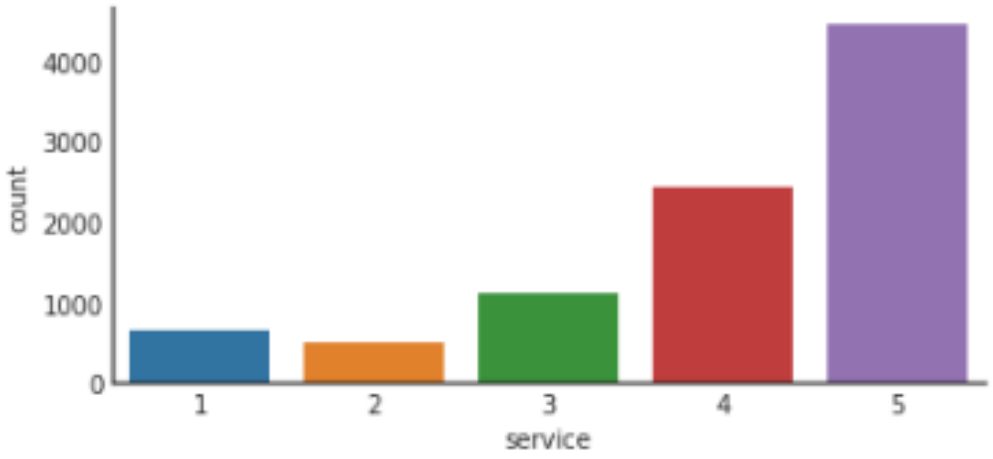


Figure 25 : Rating for 4th restaurant

CHAPTER 7: CONCLUSION

5.1) OUR INFERENCE:

After analyzing data for 30 restaurants we observed that by using natural language processing and machine learning algorithms we can process the online reviews and produce data in an easy to read manner, with reasonable accuracy. By studying seven research papers we concluded that individual aspect analysis of textual data is much more accurate than analysis of numeric rating provided by user online. We used python NLTK for data analysis and scikit for regression analysis to check accuracy of our results. Sill a lot of data processing needs to be done to produce a much more efficient and accurate sentiment analysis, of online reviews of restaurants.

5.2) FUTURE SCOPE:

In future I hope to extend working of this project to various other domains like analyzing hospital data to infer which hospital specializes in which domain, similarly analyzing e-commerce sector. There are many other fields in which this sentiment analysis concept can be applied. Finding which schools are best in a city, divided on the basis of children of different age groups. Which colleges are best for science, commerce and humanities depending on public reviews.

REFERENCES

[1] Gilbert, CJ Hutto Eric. "Vader: A parsimonious rule-based model for sentiment analysis of social media text." *Eighth International Conference on Weblogs and Social Media (ICWSM-14). Available at (20/04/16) http://comp. social. gatech. edu/papers/icwsm14. vader. hutto.pdf*. 2014.

[2] Gupta, Narendra, Giuseppe Di Fabbrizio, and Patrick Haffner. "Capturing the stars: predicting ratings for service and product reviews." *Proceedings of the NAACL HLT 2010 workshop on semantic search*. Association for Computational Linguistics, 2010.

[3] Ganu, Gayatree, Noemie Elhadad, and Amélie Marian. "Beyond the stars: improving rating predictions using review text content." *WebDB*. Vol. 9. 2009.

[4] Blair-Goldensohn, Sasha, et al. "Building a sentiment summarizer for local service reviews." *WWW workshop on NLP in the information explosion era*. Vol. 14. 2008.

[5] Kim, Soo-Min, and Eduard Hovy. "Automatic identification of pro and con reasons in online reviews." *Proceedings of the COLING/ACL on Main conference poster sessions*. Association for Computational Linguistics, 2006.

[6] Mullen, Tony, and Nigel Collier. "Sentiment analysis using support vector machines with diverse information sources." *Proceedings of the 2004 conference on empirical methods in natural language processing*. 2004.

[7] Nasukawa, Tetsuya, and Jeonghee Yi. "Sentiment analysis: Capturing favorability using natural language processing." *Proceedings of the 2nd international conference on Knowledge capture*. ACM, 2003.

APPENDICES

An introduction to machine learning by the community Digital Ocean referred on 28 November, 2017.

Introduction to machine learning tutorials by all programming tutorials referred on 28 November, 2017.

Sentiment technology by lexanalytics referred on 28 November, 2017.

A paper on Multi-document-English-Text-Summarization-using-Latent by ijser.org, referred on 28 November, 2017.