LONDON METROPOLITAN UNIVERSITY

islington college
(इस्लिङ्टन कलेज)

FINAL YEAR PROJECT PROPOSAL

Sentiment Analysis of Hotel Reviews mined from Trip Advisor

2019-20 Autumn

Student Name	**:**	**Riya Shakya**
London Met ID	**:**	**17031225**
College ID	**:**	**np01cp4a170134**
Assignment Due Date	**:**	**November 22nd, 2019**
Assignment Submission Date	**:**	**November 22nd, 2019**
Word Count	**:**	**2400 words**
External Supervisor	**:**	**Ishwor Shrestha**
Internal Supervisor	**:**	**Subeksha Shrestha**

I confirm that I understand my coursework needs to be submitted online via Google Classroom under the relevant module page before the deadline for my assignment to be accepted and marked. I am fully aware that late submissions will be treated as non-submission and a mark of zero will be awarded.

Table of Contents

List of Figure:

List of Table:

1. Introduction

a. Problem scenario:

Hotels are an imperative part of the tourism industry. The success of a hospitality business like a hotel depends on the client whether the client enjoys the hotel's services or not. The consumers' opinions on hotel facilities can be known through customer reviews. Trip Advisor is one of the largest online travel review site with a strong network effect, where guest reviews can be found. It influences 40-50% of all online travel with an annual growth rate of 43%. (Advisor, 2019) The reviews of Trip Advisor greatly impact the business as it is the foundation to rating, influence booking, and evolvement of business all over. Reviews provide a strong value for the hospitality business as more reviews is more engagement and the mistake made by many hospitality businesses is not actively collecting guest reviews on sites like Trip Advisor. Trip Advisor has more than 280 traveler reviews and opinion submitted to the site per minute which makes it difficult to go through the hotel reviews manually (Advisor, 2019). Therefore, it is hard for hotel business owner to determine what customer loves and hates based on huge data of reviews.

b. Project as a solution:

A web application will be developed to tackle the problem mentioned above. The application will have an input area where the user can input Trip Advisor's hotel link. The web application will analyze the reviews of Trip Advisor and give proper sentiment analysis of reviews in form of textual and virtualization. The website will help the company to know whether the product is getting good feedbacks from the consumers or there is a weakness that needs to change or perhaps marketing policy is not practical and many other factors.

2. Aims and Objectives

The aims of this project are listed below:

- To gain knowledge about Natural Language Processing (NLP) and Machine Learning (ML) to create a web application.
- The NLP will be an effective tool to build a foundation for Part of Speech (POS) tagging and sentiment analysis.
- ML techniques will help to solve complex natural language processing tasks, such as understanding double meaning through automated training.
- The reviews of the customers are mainly unstructured which are difficult, time-consuming and expensive to analyze, understand and sort through but this project aims to make sense of the unstructured text by automating business processes, getting actionable insights, and saving hours of manual data.

The objective of the project is

- To develop a web application which will reflect the learning outcome of NLP and ML.
- To give a textual and virtual presentation of the reviews of Trip Advisor through the application will give which will give the hotel owner a clear idea of the customer review.
- The project will deliver a structured review of the clients to the hospitality business for the improvement of the hotel.

3. Expected Outcomes and Deliverables

- The expected outcome is the development of a web application which will deliver the textual and virtual presentation of the reviews of guest in Trip Advisor of the hotel.
- The application will list out the most positive and negation feedback of the hotel from the countless number of reviews mined from Trip Advisor with the help of Sentiment Analyses.

- The hotels can refer to this project for improvement of the hotel as it points out the negative reviews which are the downfall for the reputation of the hotel.

4. Project risks, threats and contingency plan

The sentiment analysis is a new type of “digital hybrid” which is an overlap of computational linguistics, NLP and text analysis, with an increasing generous dose of AI blended into the mix. It has the power to shred light in issues which is difficult to measure or simply ignored till now. (Mccollum, 2016) The new concept is not perfect yet. This project has risks as there is limitation to Sentiment Analysis which are listed below:

- It has automation issue as the algorithm would not be able to perceive the “tone” in the writer’s voice. It faces difficulties to identify and parse humor and sarcasm in text for automated sentiment analysis platform.
- It has only 65 to 70% accuracy even in the best of circumstances. The accuracy rate drops even further when the process is applied text in any other language than English. (Heires, 2015)
- It has ethical issues as it is legal in some countries like the United States but illegal in the European Union.
- It has internal threats and instances of data loss.
- The data extracted could be bias.

The possible risks and threats of this project can be dealt by:

- The automation issues need to recommend a proper sentiment library to correctly determine sentiments and scores in words and phrases.
- The proper analysis of a sentence for sentiment will require to break down sentences into pieces including Part of Speech POS-tagging.
- The ethical issues can be dealt with proper follow up of ethical rules.
- The internal threats and instances of data loss will require data back-up to be on the safer side.
- To prevent bias results, the model should be well-trained.

5. Methodology

The methodology which will be used for the development of the web application is Iterative model. It is an implementation of a software development life cycle (SDLC) that focuses on initial and simplified implementation, then further progresses to gains more complexity and border feature set until the completion of the system. The methodology is based on the concept of incremental development, which is often used liberally and interchangeably, it explains the incremental alterations made during the design and implementation of each new iteration. (Powell-Morse, 2016)

The first step is the extraction of data from Trip advisor which will create a list of reviews. The next step is data processing which is to convert the text in lower case and remove punctuation. The next step is tokenization which will create a vocab to Int mapping dictionary, encode the words and encode the labels. The next step is training, validation, test dataset split, the next step is data loaders and batching. The next step is defining the model class. The last step is testing on test data or user-generated data. The reason to select iterative methodology for this project is the repetition of the filtering for better performance as there are terms as bigrams and trigrams which plays a vital role in the sentiment score as the result of this project.

6. Resource Requirements

The resource required for the completion of the project is a hardware device - laptop, and software which includes Python, Python framework Django and Trip Advisor. The reviews for hotels will be extracted from trip advisor using python (web scraping). The mined data will go through sentiment analysis which uses evaluation metrics for classification problem. The python framework Django with help with the textual and virtual presentation.

7. Work breakdown structure

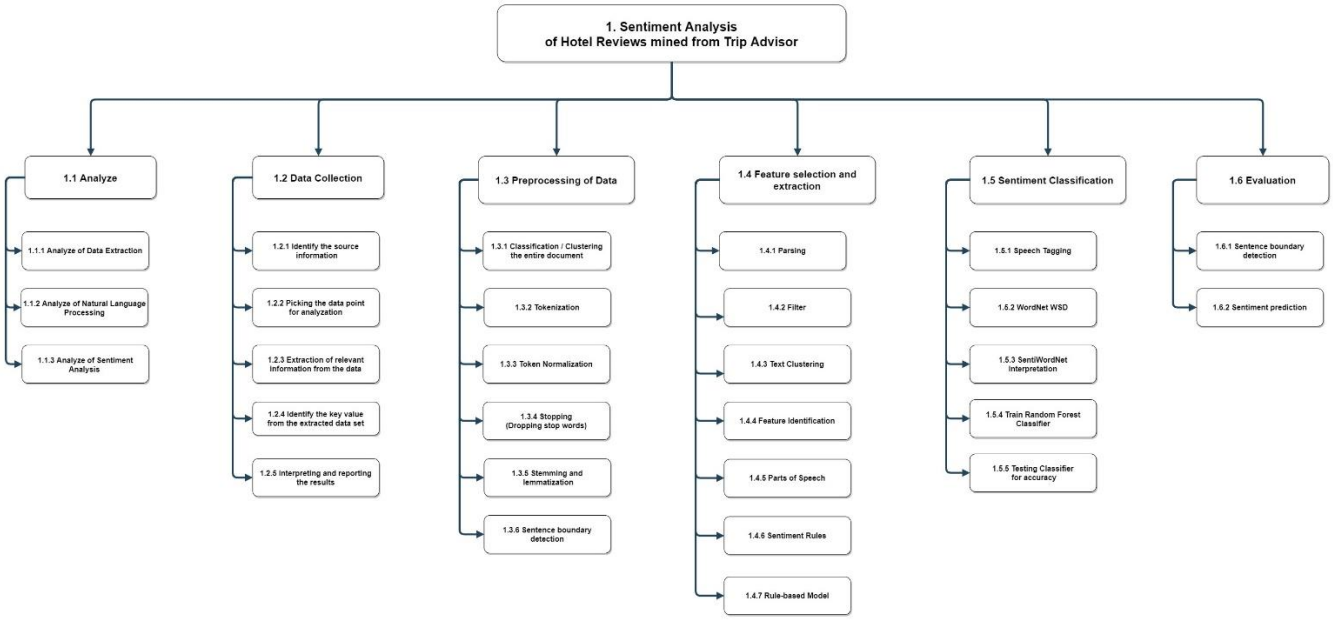

Figure 1. Work breakdown structure of Sentiment Analysis of Hotel Reviews mined from Trip Advisor.

Project Category	Description	Duration (In days)
Analyze of Date Extraction	The process of data extraction from website was studied through various research.	4
Analyze of Natural Language Processing	The general knowledge of Natural Language Processing (NLP) was gained through research.	5
Analyze of Sentiment Analysis	The different phases of sentiment analysis were studied thoroughly to understand the workflow of sentiment understanding.	4
Identify the source information	Search for the best way to extract data.	3
Picking the data point for analyzation	Listing out the websites and data which is to be extracted.	2
Extraction of relevant information from the data	The process of data extraction is implemented.	13
Identify the key value from the extracted data set	The extracted data is filtered out, so we have only the required data(reviews.)	5
Interpreting and reporting the results	The filtered-out data is arranged properly.	3
Classification / Clustering the entire documentation	The extracted data goes through the process of classification or clustering.	5
Tokenization	The sentence is broken down into words.	4
Token Normalization	The process of canonicalizing tokens so that matches occur despite their difference in the character sequences of the token.	5
Stopping (Dropping Stop words)	The stop words such as “the”, “a”, and so on are removed.	2
Sentence boundary detection	The process where NLP decides where a sentence start and ends.	5

Parsing	The process of determining the syntactic structure of text by analyzing its constituent words based on underlying grammar.	4
Filter	The process to identify the meaning of the words.	3
Text Clustering	The process to understand and categorize unstructured, textual data.	9
Feature Identification	The process conducts word counts, stop word counts, punctuation counts, the length of characters, the language of text, vectorization, stemming, part of speech tagging, and the name of the entity extraction and many more.	18
Part of Speech	The process read text in one language and assigns part of speech to each word such as noun, verb, pronoun and so on.	15
Sentiment Rules	The process to identify an d extract opinions a text.	5
Rule-based Model	The process to understand, analyses and manipulate general language.	3
WordNet WSD	The process consists of associating words in context with their most suitable entry in a pre-defined sense inventory	7
SentiWordNet Interpretation	The process is like WordNet with determining the sentiment of words.	15
Train Random Forest Classifier	The process to train the model.	30
Testing Classifier for accuracy	The process to test the train model.	10
Sentiment prediction	The process where the sentiment of the words.	5

Table 1. Work Break Down structure table

8. Milestone

The milestones for this project are listed below:

- **Data Acquisition**: The data mined from Trip advisor is important as it is required for analyzing and classifying text in the dataset. (December 11th, 2019)
- **Text preprocessing:** The preprocessing phase is required to reduce noise in data by removing unnecessary stop words, repeated words, stemming, removal of emotion, removal of URLs, and such. (December 16th, 2019)
- **Feature selection and extraction:** The proper selection and extraction of features which is key in determining the accuracy of the model. (January 3rd, 2020)
- **Sentiment Classification:** The sentiment classification techniques such as Naïve Bayes and Support Vector Machines are applied to classify the text. (January 30th, 2020)
- **Polarity detection:** The determination of polarity of the sentiment to detect whether the test is positive, negative or neutral. (March 3rd, 2020)
- **Validation and evaluation:** The overall accuracy of the techniques of sentiment analysis is determined through validation and evaluation of the obtained results. (Dobrescu, 2011) (March 9th, 2020)

9. Project Gantt chart

	Task Name	Duratio	Start	Finish
1	**Phase 1: Iteration 1**			
2	**Analyze**			
3	Analyze of Data Extraction	6 days	Tue 11/5/19	Tue 11/12/19
4	Analyze of NLP	5 days	Thu 11/7/19	Wed 11/13/19
5	Analyze of Sentiment Analysis	9 days	Sun 11/10/19	Wed 11/20/19
6	**Requirements**			
7	Data Acquistition	15 days	Thu 11/21/19	Wed 12/11/19
8	Text preprocessing	7 days	Sat 12/7/19	Mon 12/16/19
9	**Analysis**			
10	Feature selection	7 days	Tue 12/17/19	Wed 12/25/19
11	Feature extraction	11 days	Sun 12/22/19	Fri 1/3/20
12	**Design**			
13	Sentiment Classification	20 days	Sat 1/4/20	Thu 1/30/20
14	**Testing**			
15	Create model	18 days	Wed 1/22/20	Fri 2/14/20
16	**Implementation**			
17	Polarity detection	18 days	Fri 2/7/20	Tue 3/3/20
18	**Review**			
19	Validation and evaluation	7 days	Sat 2/29/20	Mon 3/9/20
20	**Phase 2: Iteration 2**			
21	**Design**			
22	Classification for bigrams	7 days	Tue 3/10/20	Wed 3/18/20
23	**Testing**			
24	Create new model with new data for bigrams	7 days	Fri 3/13/20	Mon 3/23/20
25	**Implementation**			
26	Polarity detection for bigrams	8 days	Mon 3/16/20	Wed 3/25/20
27	**Review**			
28	Validation and evaluation	6 days	Sun 3/22/20	Fri 3/27/20
29	**Phase 3: Iteration 3**			
30	**Design**			
31	Classification for trigrams	10 days	Sat 3/28/20	Thu 4/9/20
32	**Testing**			
33	Create model with new data for trigrams	6 days	Mon 3/30/20	Mon 4/6/20
34	**Implementation**			
35	Polarity detection for trigrams	8 days	Wed 4/1/20	Fri 4/10/20
36	**Review**			
37	Integration and Test Support	8 days	Sun 4/5/20	Tue 4/14/20
38	**Deployment**			
39	Plan Deployment	6 days	Wed 4/15/20	Wed 4/22/20
40	Develop Support Material	7 days	Sat 4/18/20	Mon 4/27/20
41	Produce Deployment Unit	7 days	Mon 4/20/20	Tue 4/28/20
42	**Maintenance**			
43	Manage Acceptance test (At Development Site)	7 days	Wed 4/22/20	Thu 4/30/20
44	Manage Acceptance test (At Installation Site)	7 days	Fri 4/24/20	Mon 5/4/20
45	**Documentation**			
46	Interim Report	30 days	Thu 11/28/19	Wed 1/8/20
47	Final Report	60 days	Wed 2/5/20	Tue 4/28/20

Figure 2 Gantt Chart of Sentiment Analysis of Hotel Reviews mined from Trip Advisor.

10. Conclusion

The project is mainly concerned with bringing a change in the field of tourism. Tourism is revenue recognition of the country. The tourist attraction does not only include scenery and adventures but also the hotel. The facilities provided by the hotel are important for the tourists. The better facilities provided by the hotel will attract more tourists. The improvement of the hotels requires reviews of the customer. The numbers of reviews of the hotel in the website will be uncountable. Hence, the sentiment analysis is used to resolve the problem related to hotels. The sentiment analysis on the mined reviews will provide highlighted reviews.

The project will save the hotel a lot of time while listing out the positive and negative reviews from the mined data. The sentiment analysis is a process of computationally identifying and categorizing opinions from a piece of text and determine whether the writer's attitude towards a topic or the product is positive, negative or neutral. The expected outcome of sentiment analysis is a list of positive reviews on the facilities provided by the hotel and the negative reviews on the facilities provided by the hotel. The negative reviews on the facilities will help the hotel to work on the facility for the betterment of the hotel. It mainly focuses on customer satisfaction.

The phase of data mining is where risk arises in the project. The data mining from a website will require good knowledge of programming and sentiment analysis phase will have numerous trials and errors. Hence, the methodology iteration is used. The resource required for this project is a good knowledge of python, ML and NLP. The work is divided into different stages for an easier flow of development of the software. The milestones are listed for the indication of the project. All in all, the completion of the project will be required a lot of studies and researches. The success of the project will not be easy to achieve, but it is also not impossible with strong determination and hard work.

Bibliography

Advisor, T. (2019) *TripAdvisor Network Effect and the Benefits of Total Engagement* [Online]. Available from: https://www.tripadvisor.com/TripAdvisorInsights/w828 [Accessed 5 October 2019].

Dobrescu, A.B. (2011) *https://rua.ua.es/dspace/bitstream/10045/19437/1/tesis_alexandrabalahur.pdf*. 7th ed. Carretera San Vicente del Raspeign s/n: Universitat d'Alacant.

Heires, K. (2015) *Sentiment Analysis: Are You Feeling Risky?* [Online]. Available from: http://www.rmmagazine.com/2015/12/01/sentiment-analysis-are-you-feeling-risky/ [Accessed 5 Ooctober 2019].

K, J. (2018) *12 BEST SOFTWARE DEVELOPMENT METHODOLOGIES WITH PROS AND CONS* [Online]. Available from: https://acodez.in/12-best-software-development-methodologies-pros-cons/#Extreme_Programming_Methodology [Accessed 9 September 2019].

Mccollum, O. (2016) *More Than Just a Gut Feeling: Using Sentiment Analysis in Risk Management* [Online]. Available from: https://www.business2community.com/social-data/just-gut-feeling-using-sentiment-analysis-risk-management-01470254 [Accessed 5 September 2018].

Negnevistky, M. (2005) *Packet*. 2nd ed. PackEdinburgh Gate: Pearson Education.

Powell-Morse, A. (2016) *Iterative Model: What Is It And When Should You Use It?* [Online]. Available from: https://airbrake.io/blog/sdlc/iterative-model [Accessed 05 September 2018].

Rothman, D. (2018) *Artificial Intelligence By Example: Develop machine intelligence from scratch using real artificial*. Birmingham Mumbai: Packet.