CHAPTER 5

OPINION MINING FROM SOCIAL CONVERSATIONS

5.1 Overview

Social media democratised communication medium ignoring all the existing norms of hierarchies, structures, gender and parameters bridging other categories of the social divide. This channel has totally changed the way people react and comment their views on any services or products. Users, post their opinion in multiple locations. These user-generated contents are the new electronic Word of Mouth (eWoM) information promoting media. The explosion of these kinds of social media sites promoted marketing. The worldwide trend of shifting the marketing from the traditional channels to social media has resulted in building a new kind of customer relationship. In any domain, there is an increasing trend to check online to get the opinion before taking a final buying decision. A product or service promoted by word of mouth has a strong personal influence published virally. Social media which started as a casual medium of communication has emerged as eWoM today targeting social, functional and emotional factors which are turned into favourable factors. Social networking platforms such as Twitter, Facebook, YouTube and LinkedIn help the connectivity promoting the

reach, as part of their broader customer engagement programs. Social media marketing is an effective way to engage audiences. This mode of marketing helped in faster information diffusion and expanded the reach of the brand by driving more traffic into the sites. Just like how users use it for decision making, the business also has the option of knowing the review of their product. It is extremely important to understand user preferences and behaviours.

5.2 Challenges

Sentiment analysis deals with text analysis. Social media expressions contain sarcasm, acronyms, jargons, slangs and emoticons which cannot be understood by the traditional technological means. The analysis should be more precise than just polarity measurement to the precision of success, curiosity etc. A word can take different meaning depending on context. The word book used as noun and verb carries different meanings. Domain dependency needs to be taken care of. Abbreviations like LOL, ROFL etc. popularly used cannot be found in dictionaries. The sentiment analysis can be at the document-level, sentence-level or entity-level.

5.3 Scenario

This is the case study of having a real-time capability for listening and analysing the opinions expressed by the customers in various channels and sites and thereby improving the business process and generating a competitive advantage for an eco-nature tourist property. This property is

located in the Wayanad District in the State of Kerala, in India. We start considering the online reviews that appeared through social media Facebook and later extended to other channels. Eco-tourism properties are, especially for the nature loving tourists. It keeps its ambience more, as an environment friendly property closer to nature, with a green environment. This promotes visitors enjoying the nature to enjoy a greener world. Here we try to discover consumer preferences about ecotourism, speciality features of the property and restaurant, using opinions available on the different review sites.

An automated tool is developed which extracts data from reviews available on social media, containing valuable information about customer preferences. The tool helps in determining the sentiment orientation of opinions and the feedback to improve the services offered which in turn increases the traffic.

The aim is to

- Reach out and create a community of similar people interested in travelling, exploring new places, cuisine and experiences.
- Involve the community and strengthen the customer engagement, connectivity and enhance customer satisfaction thereby improve retention rates

- Develop a reputation management plan to respond to online review comments
- Target the segmented areas effectively by promoting community members to generate content and share their experiences and expectations
- Provide real-time service to the customers
- Understand the trends and tipping points in customer patterns

5.3.1 Developing a Response Review Strategy

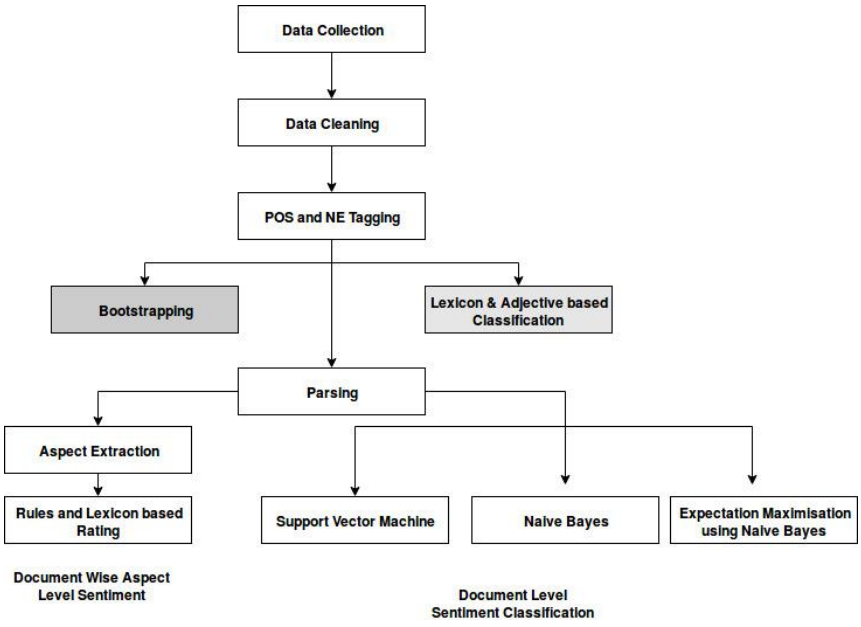
Customers share their views about the resort on the web in the form of reviews, blogs, comments, etc. A customer review is a user generated content about the property. People interested in visiting such a location will be looking for such reviews. Probable customers read this review. This can be influential in the decision making process of choosing the destination. A large amount of information is available but at varied locations. The difficulty due to the information overload makes the user difficult to read all and decide. Hence the automatic extraction of the data and presenting to them the relevant information is very important, from the business angle. Many aggregators make use of these functions. Social media has to be constantly monitored to listen to the mention of your product, as well as your competitor's product that is relevant to your business.

There are multiple data sources available. Here, the study starts with social media communications through Facebook. Some of them are very straightforward expressions. Most of the times, it may not be explicit conversations. Aggregator sites like TripAdvisor, Agoda, Goibio, etc. are increasingly used by people for review as well as decision making. Service industry like Tourism has multiple factors to be considered, which impact customer sentiment and opinions. The opinion data is temporal and dynamic. The opinion mining tool will identify and extract subjective information to determine the sentiments expressed. Corrective action is to be taken if there are any negative comments. We expect the customers to rate the property with reference to some standard facets like location, cleanliness, service, food etc. But it is seen that weather, health directive, approach road, noise from construction neighbourhood and perception also comes in the review.

In this case study, from the user's key behavioural pattern, the following things are analysed

1. What people talked about their experience in the property?
2. What do they like/dislike about the property?
3. What are they talking about you?
4. How many people are talking about you?
5. How to capitalise the influence?
6. Is there any demographic usage pattern?

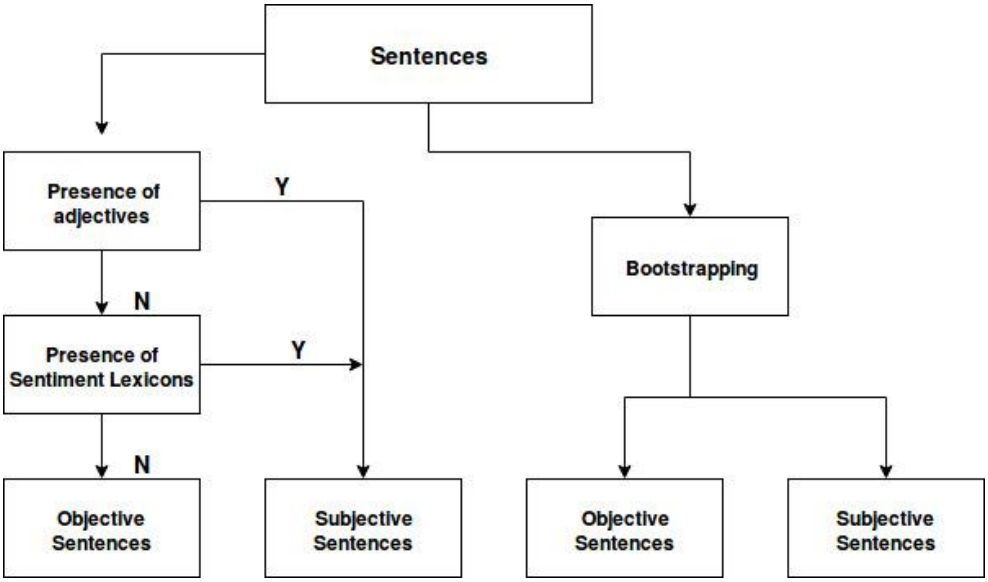
For analysing the conversations, the various classification techniques used are discussed in chapter 3. The detailed workflow is also given in the chapter. In our case study, we use a hybrid algorithm applying both supervised and unsupervised classification techniques at different steps of the process. Figure 5.1 shows the on steps involved in applying document-level or aspect-level classification algorithm.



Source https://www.linkedin.com/pulse/sentiment-analytics-trilogy-master-class-extracting-mukherjee
Fig 5.1 Flow diagram of the procedure of Sentiment Classification

Sentences are divided into subjective and objective. Subjective sentences express information as opinions, speculation, judgment and similar. Objective sentences are factual representation. For improving the efficiency of the algorithm, here objective sentences are discarded. On these classified

subjective sentences, document-level or aspect-level classification algorithm is applied based on the requirement. Document-level classification gives the overall sentiment, but aspect-level is needed for knowing where improvement is needed. Figure 5.2 shows how the sentences are classified by checking the contents of the sentence.



Source: https://www.linkedin.com/pulse/sentiment-analytics-trilogy-master-class-extracting-mukherjee

Fig 5.2 Sentence Classification

5.4 SENTIMATCH -The Tool

SENTIMATCH tool is modelled and developed to understand opinions, emotions, sentiment in the contextual level beyond keyword matching. Customer engagements are captured which help in effectively segmenting

the targets like geographical targeting or demographic targeting. A user leaves behind an information trail when a website is visited. These semi-structured website log files are captured as Clickstream data. The data elements including the visitor's IP address, the URLs of the pages visited, and a user ID that uniquely identifies the user, time stamp are available in the log files. Clickpath or Clickstream is the route the visitors navigated.

SENTIMATCH constantly monitors the conversations by continuously crawling and indexing sites. They look for new customers as well as the returning customers. Attention needs to be paid to trending negative comments, in positive reviews as well. This added to machine learning intelligence helps in understanding customer perception, and improves online reputation.

SENTIMATCH extracts all opinions from the social media, summarises the review and presents an overall perspective. The feedback helps in improving the services of the property.

Sentiment can be analysed at different structural levels in the text ranging from individual words to the entire documents. Liu [101] has defined classification under Sentence-Level, Document-level, and Aspect-Level sentiment analysis.

The default or baseline analysis scans the entire documents. A sentiment score is calculated for the whole piece of text submitted for analysis. A document can contain multiple paragraphs or even one sentence. In the case

of sentence level analysis, a detailed analysis is done, and a sentiment score is generated for each sentence. This is not an average of the entities or words in each sentence but an exhaustive aggregation of each entity in the sentence through exhaustive grammatical analysis. Entity-level Sentiment is a granular examination of sentiment contexts and signals in the text**.**

To understand this, consider the following example of a Facebook post

"Mesmerizing Place!!! Natural beauty and mud (earth) rooms are top notch. To be honest, i was not expecting much when viewed the photos online and did the booking, but woow i was in for a surprise when i checked in."

Level of Classification	Sentiment
Document	Positive

Fig 5.3 Document-Level Classification

Level of Classification	Sentence	Sentiment	Sentiment Score
Sentence	Mesmerizing Place!!!	Positive	0.7
	Natural beauty and mud (earth) rooms are top notch	Positive	0.5
	To be honest i was not expecting much when viewed the photos online and did the booking , but woow i was in for a surprise when i checked in .	Positive	0.9

Fig 5.4 Sentence-Level Classification

Document-level classification of the given document is shown in figure 5.3 . In the sentence level classification, figure 5.4 the sentiment score is also included. The package RSentiment analyses sentiment of a sentence in English and assigns score to it. The function calculate_sentiment gives the sentiment score. Sentences are classified into categories: positive, negative, very positive, very negative and neutral. The sentiment function returns a (polarity, subjectivity) tuple, for the given sentence, based on the adjectives it contains, where polarity is defined as a value between -1.0 and +1.0 and subjectivity between 0.0 and 1.0. This may also be extended to entity-level classification identifying each word in the sentence.

5. 5 Experiment set up- Algorithm

We use a hybrid algorithm combing supervised and unsupervised classification techniques at different steps of the procedure.

The steps involved are

- Acquire data by crawling and scraping dynamic web pages and parse online reviews
- Pre-process data and convert it into a usable form, cleaned and ready for analysis
- Classify the data at document or aspect level
- Run and train models for opinion analysis
- Provide the final result for analysis.

5.5.1 Software Set up

Python framework: Django

Python package: pymongo, Numpy, Scipy, NLTK, VADER, PyTorch

PyTorch provides a high level feature for Tensor computation (like Numpy) with strong GPU acceleration.

Database: MongoDB, ELASTIC

Apache Spark Clusters are used for fast computation. Apache Solr enables the fast and easy creation of search engine environment. Solr can be treated as sentiment analysis and retrieval tool. The UIMA (Unstructured Information Management Architecture) framework and multi-language analysers are used for sentiment extraction. The sentences are passed through pluggable annotators which detects Entity and its associated polarity. Solr indexes stores polarity of each sentence. Persistent model files can be created from training data and accessed at runtime. Lucene, Solr provides fast, highly scalable and easily maintainable full-text search capabilities. Solr can be considered as a sophisticated token-matching engine.

Python is the programming language chosen, as it is found suited for processing large data sets and performing mathematical computations. It has many libraries that support machine learning and textual analytics. Natural Language ToolKit (NLTK) is one of the Python packages used here for sentiment analysis. Apart from Python, R scripts are used for utilising processing results of text mining libraries. Beautiful Soup, another python library creates a parse tree from parsed HTML and XML documents (including documents with non-closed tags or tag soup and other malformed mark-up's) are also extensively used. JavaScript (JS) is more used for metadata adding and hence good for data mangling and data de-duplication.

Using Python with lxml and Requests permits web scraping easily. Scrapy is a Python framework for ETL (Extract, transform and load) to build a

customized crawler, parser, data scraper and converter for extracting structured data from websites. Scrapy is relatively simple, flexible and very powerful.

NLTK package is used to build a sentiment analysis model on the prepared dataset. Preparation is explained in the later part. NLTK classifiers work with feature structures. This is a simple mapping of dictionary feature name to a feature value. We initially used the Naive Bayes Classifier, which made predictions based on the word frequencies associated with each label of positive or negative. But it was found that Naive Bayes doesn't consider the relationship between words. It could not recognise the fact that "not" is a negator to the word "bad". Instead, it read and classified two negative indicators as such.

Apache Spark Clusters are used for fast computation. Apache Solr enables the fast and easy creation of the search engine environment and also helps in sentiment analysis. The UIMA (Unstructured Information Management Architecture) framework and multi-language analysers can be utilised for sentiment extraction. NLTK has a built-in method that computes the accuracy. A python script is run to find the accuracy of our classifier
The library VADER (Valence Aware Dictionary and sEntiment Reasoner) [107] is a lexicon and rule-based sentiment analysis library that works well on a short text as well. VADER helps in performing sentiment classification quickly even if train data is small and can write custom code to search for words in a sentiment lexicon. Hence it gives good results for

sentiments expressed in social media.

For sentiment analysis with VADER, we need to define ranges for compound value to categorize the sentiment of the document as extremely negative, negative, neutral, positive and extremely positive. The VADER algorithm outputs sentiment scores to 5 classes of sentiments. When you run the sentiment analyser on a document, varying sentiment proportions for individual sentences are obtained as shown below:

Great place to be when you enjoy Nature!			
negative : 0.0	neutral : 0.702	compound: 0.6249	positive: 0.412
The food is really GOOD!			
negative : 0.0	neutral : 0.4901	compound: 0.6421	positive: 0.509
The food is good but not all that great.			
negative : 0.172	neutral : 0.623	compound: 0.1204	positive: 0.205

Table 5.1 Sentiment Scores

The compound value here conveys the overall positive or negative user experience. There is a difference in compound sentiment in the above three instances given. The sentiments expressed both before and after the 'but' are taken into consideration. Hence there is a change in a sentence's sentiment intensity.

5.5.2 Model Framework for feature extraction

The eco-tourist property is interested in knowing, what other people are talking about their property, service or any other issues affecting them. This property gets about 350 to 600 review pages annually. Here we are using feature-based sentiment analysis. The various modules include data acquisition, feature extraction, sentiment prediction, and classification. The tool developed focus on polarity finding, and hence the sentiment expressed by the extraction of customer opinions from unstructured reviews is provided at various review sites.

The process involved is explained in the steps below

5.5.3 Data acquisition and Pre- Processing

People use social media platforms to express their opinion directly. The system constantly monitors the social media conversations. The data acquisition is from different locations. Data are from different sources and hence may be heterogeneous in nature including text, language and domain. The data pointing to or mentioning the ecotourism property is collected from various sources including, Facebook posts, reviews from TripAdvisor, Agoda, Goibibo, tweets, media reports, and volunteer messages. The input to the opinion mining process is collected set of such reviews.

The Acquisition module takes care of data pre-processing. The data is unstructured and may be of different formats. There may be overlapping

data also. Only English language conversations are considered. All irrelevant facts are discarded.

Data cleansing include reformatting, de-duping, merging, and filtering. Data cleansing is basically the scrubbing and cleaning applied to raw data and getting it into a format required for analysis. A data is said to be tidy if the table has variables in columns and observations in rows. The `dplyr` and `tidyr` packages of R are used to optimise the data wrangling process. Pre-processing improves the accuracy of data by avoiding unwanted data processing in each phase. The words like Oh!, OMG!, emotions and other alphanumeric characters are removed. The concept of the quintuple structure of Bing [66] is followed

5.5.4 Data Curation

Data is acquired from different sources like owned media and earned media. The property owners have no control on earned media, where there is a mention about the property. There is a critical need for sharing the context of data created. This is called data curation. Data curation focus on how semantics can be added for context-sensitive usage of data. The learning is by tracking the social network, and social interactions between users of data. Curation is about collecting, tagging, annotating data and governance to ease out the process from the beginning onwards. The content acquired and cleansed needs an additional layer of segregation and normalisation for further processing. There can be data overlap also. The challenges here are the data format compatibility. A tweet data and an FB post cannot be joined

together even if they have related information because of the format compatibility. Another challenge is the photo linking to a keyword. Curation prepares the data and the metadata for ingestion and further analysis.

The most important part of the analysis is the ontology to extract weighted meta-tags. The ontology contains concepts, relationships between those concepts, and rules about how they fit together, to the multiple reviews crawled and stored. The ontology extracts meaning from structured or semi-structured data. We need to identify people looking for "Wayanad, Eco tourism, Kerala hill resort, Mud structure, Pazhassi Raja, Banasura Sagar etc..." Example of ontology is as follows: The destinations fit into vacation themes — like a Family vacation, Outbound training, Plantation Stay, Corporate retreat etc.

The SENTIMATCH will do the opinion mining to see if the user liked the property or the activities happening there. This is fed to rule engine. The rule engine looks for sentiment words such as "like", "love", "ambience," "hate," or "good view," and relate to what they mean and to the resort and reaction of the customer. Weightage table is created. The collected reviews are stored in a database which is used for the Opinion Mining process. From the extracted data, the first step is keyword tagging. Keywords are matched against the taxonomy (based on weights allotted). Keywords that do not match are listed. Taxonomy is evolving. Once the taxonomy is created, the taxonomy software uses natural language processing, semantic analysis, and

statistical pattern matching. The body of text is analysed and then assigned the taxonomy. A metadata tag is linked. There can be a modification or overriding of the resulting classification. Manual intervention may be needed at times for ambiguous documents. Ontology is an explicit specification of the description of the area of interest, which matches the end-users requirements. The application of ontologies and semantic technologies allow the inference. In our case, even anyone looking for "Nilgiri Biosphere" also should be attracted to our site, even if there is no mention about it. The rule inference engine learns from the searching pattern, and re-ranks search results more closely fitting to the end user requirements.

5.5.5 Subjectivity Classification

Sentences as classified as subjective or objective, based on the content. Subjective sentences express some opinion whereas the objective sentences are facts. The documents obtained after data cleaning is the input for this step.

Each sentence is split into words through tokenisation. The steps followed are

1) On the basis of a list of ontology and taxonomy created, the sentences are classified as subjective and objective. Others are unlabelled.

2) These sentences are fed into the rule engine. With defined ontology

and taxonomy, lexicon based extraction pattern for subjective expressions are derived.

3) Patterns are learned continuously, which are used to identify more subjective sentences from the unlabelled set. This iterative process repeats till it cannot proceed further.

Precision and Recall are calculated to check the accuracy of the tool.

Number of Sentences Considered	80
Number of Subjective Sentences	30
Number of Objective Sentences	50
Number of Subjective Sentences classified by SENTIMATCH	24
Number of correct Subjective Sentences in the above	19

After the training phase, Facebook posts numbering 242, was subject to classification, where we obtained a precision of 88.76% and recall of 72.3%.

Precision (19/24)	**79%**
Recall (19/30)	**63%**
Precision FB Posts (213/242)	**88.76%**
Recall FB Posts (175/242)	**72.3%**

Table 5.4 Precision and Recall of Facebook Posts

5.5.6 Tokenising

The tokeniser accepts the plain text as input and a language parameter. So the entire file is broken into sentences. The input text is split into paragraphs, sentences, and tokens and generates a token with this information. The UI for accepting the plain text is given in figure 5.5

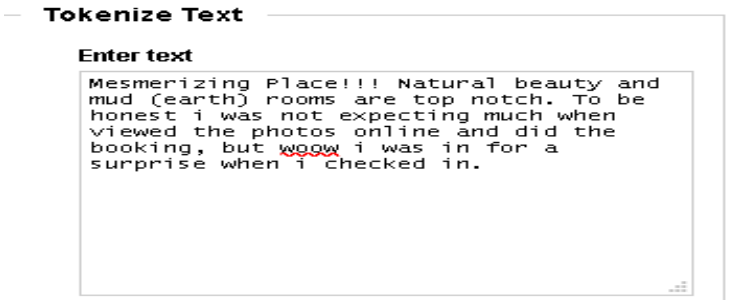


Fig 5.5 UI for Converting Text into Tokens

Tokenizer gets the list of words in strings. The output is shown in figure 5.6

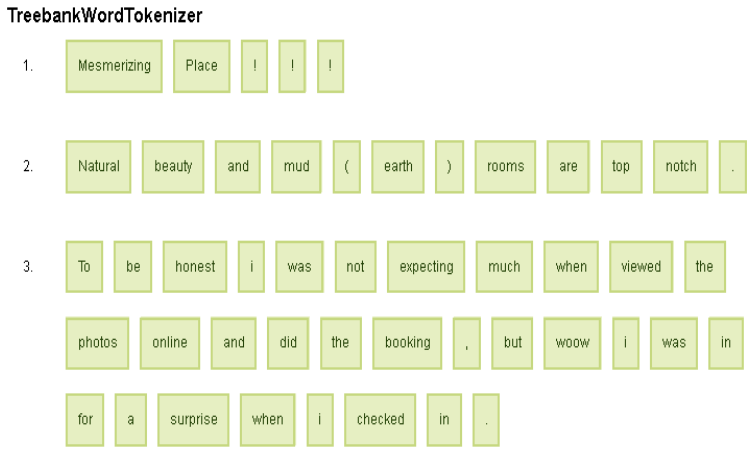


Fig 5.6 Tokenizer

The above tokens classified into their part-of-speech and tagged.

5.5.7 POS Tagger

Tagging is the process of assigning a POS marker to the words in the text. Part Of Speech (POS) tagging is a linguistic technique used for feature extraction. POS tagging assigns a tag to each word in the sentence morphologically categorising as noun, verb, adjective etc.

The default part of speech tagger is a classifier based tagger trained on the PENN Treebank corpus. The PENN Treebank corpus is composed of news articles from the Reuters newswire. That means the tagger is more likely to be correct on the text that looks like a news article, and less accurate on the text that doesn't. All the other taggers have been trained on part-of-speech tagged NLTK corpora. The output of POS tagging is shown in figure 5.7.

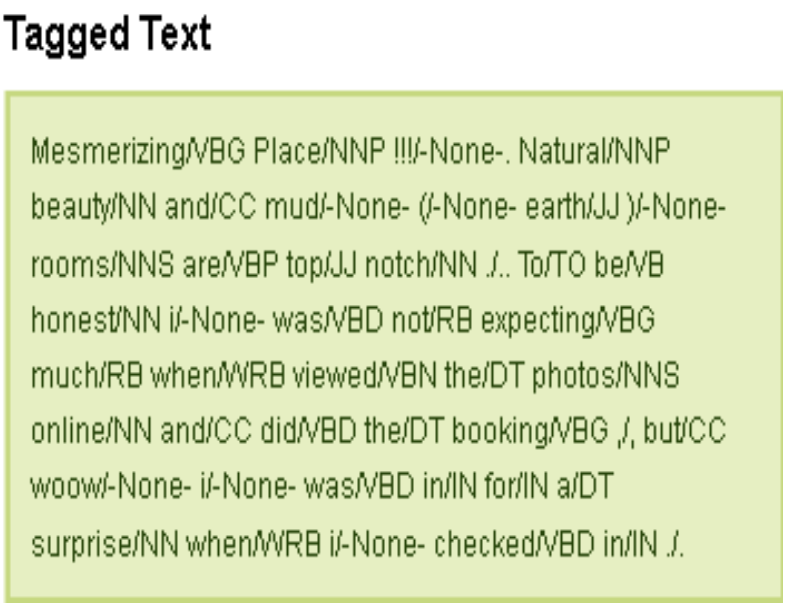


Fig 5.7 Tagged Text

All the top-level noun phrases and their associated named entity tags are identified. Chunking is the process of segmenting and labelling multi token sequences. NP-chunking is identifying chunks corresponding to individual noun phrases. Chunk grammar rules define how sentences can be chunked. A sample rule used in our chunk parser can be quoted as whenever an optional determiner (DT) followed by any number of adjectives (JJ) and then a noun (NN) is found, NP chunk is formed.

This can be visualised in figure 5.8 as given below.

Phrases and Named Entities

NP:

Place/NNP

NP:

Natural/NNP beauty/NN

NP:

earth/JJ)/-None- rooms/NNS

NP:

top/JJ notch/NN

NP:

i/-None-

NP:

the/DT photos/NNS online/NN

NP:

the/DT booking/VBG

NP:

i/-None-

NP:

a/DT surprise/NN

NP:

i/-None-

Fig 5.8 Phrases and Named Entities

5.5.8 Parsing

Parser provides syntactic tree representation of a sentence. A language

agnostic knowledge graph which determines connections/relations between the various nodes in the graph is built from the parse tree.

```
# Semantic graph: [expecting/VBG
#                         dep:[Mesmerizing/VBG
#                             dobj:[beauty/NN
#                                 nn:Place/NNP
#                                 nn:!!!/NNP
#                                 nn:Natural/NNP
#                                 cc:and/CC
#                                 conj:[rooms/NNS nn:mud/NN appos:earth/NN]
#                                 rcmod:[top/JJ
#                                     cop:are/VBP
#                                     dep:[notch/NNP infmod:[honest/JJ aux:To/TO cop:be/VB]]]]]
#                         nsubj:i/FW
#                         aux:was/VBD
#                         neg:not/RB
#                         advcl:[viewed/VBD
#                              advmod:[when/WRB advmod:much/RB]
#                              dobj:[photos/NNS det:the/DT]
#                              prep:online/NN
#                              cc:and/CC
#                              conj:[did/VBD dobj:[booking/NN det:the/DT]]]
#                         cc:but/CC
#                         conj:[was/VBD
#                             nsubj:[i/NNS nn:woow/NN]
#                             prep:[for/IN
#                                 dep:in/IN
#                                 pobj:[surprise/NN
#                                     det:a/DT
#                                     dep:[checked/VBD advmod:when/WRB nsubj:i/FW prt:in/RP]]]]]
```

Fig 5.9 Parsed Sentence

A parse tree is an ordered representation of a sentence broken into to parts-of-speech. After tokenizing, the binary relations representing the syntactic structure of a sentence are represented by dependency parse. Here it can be seen that a verb is linked to its dependents (arguments/modifiers). These relations together form a tree or tree-like graph.

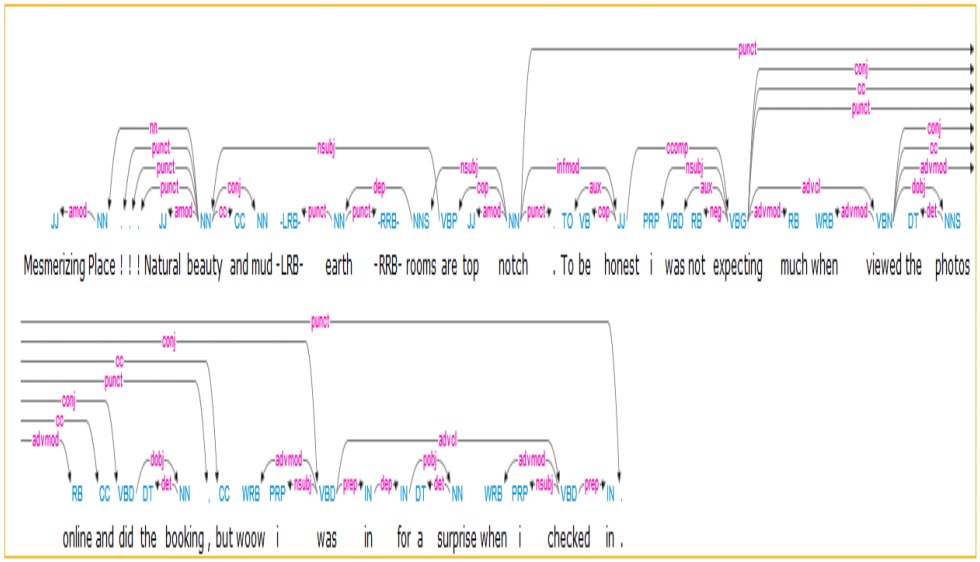


Fig 5.10 Parse Tree Structure

The above example parse tree represents the relationship between the labels which includes nominal subject as nsubj, punctuation as punct, direct object as dobj, auxiliary verb as aux, noun compound modifier as nn, determiner as det, and the prepositional phrase as prep and similar in tune with Stanford dependencies.

5.6 Sentiment Analysis and Opinion Mining

The opinion expressed can be classified in the first step as positive, negative

or neutral based on the content polarity of the sentiments expressed. Mining the contents of social media conversation- Customers perceptions are expressed in the social media through their reviews.

5.6.1 Named-Entity Recognition

People, companies, organizations, cities, geographic features, and other entities are extracted from the content, and optionally detect the sentiment of each entity. Named Entity Recognition (NER) and Classification identifies names of persons, places, etc. and classifies them in a semantic class (PERSON, LOCATION, etc.). Apache OpenNLP API is used for training model. It determines important keywords in the text, ranks them by relevance, and optionally detects the sentiment of each keyword.

Stanford CoreNLP provides a set of human language technology tools. It can give the base forms of words, their parts of speech, whether they are names of companies, people, etc., normalize dates, times, and numeric quantities, mark up the structure of sentences in terms of phrases and syntactic dependencies, indicate which noun phrases refer to the same entities, indicate sentiment, extract particular or open-class relations between entity mentions, get the quotes people said, etc.

Stanford Core NLP is a linguistic analysis tool which is part-of-speech (POS) tagger, the named entity recognizer, the parser, the co-reference resolution system, sentiment analysis, bootstrapped pattern learning, and the open information extraction tools.

The application will train different data set and materials collected by the training module. Lemmatisation is done based on dictionary-based lemmatisation, identifying lemma (dictionary entry) for the word.

5.6.2 Identifying Opinions with feature selection.

Important features are rated by the reviewers. Words which express opinion towards features are called opinion words. In the feature-based opinion mining, the features are identified. The opinion words are mostly adjectives, verbs, adverb adjective, and adverb verb combinations. For each opinion word, semantic orientation is identified. This facilitates the prediction of the semantic orientation of each sentence inputted. Care is taken in the case of negative orientation, by getting the contextual information of a sentence.

5.6.3 Rule Engine

Traditional rule-based approaches are complex. Hence statistical machine learning algorithms are used to create a powerful abstraction to the logic. Different conditions are validated. These conditions are checked against boolean (y/n) conditions, number or a character sequence called rules. The Rule engine is a framework for evaluating one or set of conditions. Rules are built. It is a continuous process. Machine learning algorithms work on it. Heuristics for improving the machine learning algorithm are required. Named entities, synonyms, topic tags, grammar and algorithms are combined with parsed keywords and logic. Different keywords are trained to find the themes and to pull out the sentiment from the word stream. Rule

engine depends on sentiment dictionaries. Rules are defined as attaching sentiments to the target or modifying sentiment scores, sentiment table, ontologies and taxonomies. The ontologies and taxonomies are also continuously evolved. Though Stanford NER is primarily used in this tool for getting sentiment information, additional NLP tools are necessary to compute and aggregate.

Rule-based classification identify words and patterns that are indicative of positive or negative sentiment like:

- polarity words: good, great, love; bad, terrible, hate
- polarity ngrams: sure shot (+), must see (+), could care less (-)
- casing: uppercase mostly indicates subjectivity
- Punctuation: Multiple punctuation marks like"!" and "?" Indicates subjectivity (often negative)

A simple Sentiment Classifier can be defined as:

Def sentiment (document) =

If (document contains "good")

positive

else if (document contains "bad")

negative

else

neutral.

Each pattern is treated as a rule; while going through the rule engine, if present in the text, the rule indicates whether the text is positive or negative.

Sometimes multiple rules may be applied which indicates both positive and negative. A compound score is calculated in that case. Rules are written in python scripts. We initially started with 22 rules. More rules were added in due course.. Currently, there are 69 rules in the rule engine. Manual intervention is required to review whether rules need to be modified.

The sample rules may be:

If a person visits through “google” and searches “eco-resort” then classify “probable”

If time spent on page =”rate” then popup “chat box”

5.6.4 Feedback

Business can be improved by an effective social media strategy which helps in sales by increasing customer loyalty. SENTIMATCH gives the most insightful content to analyse. The reviews help the business to know what their guests feel about the property. This can be considered as a reputation management program, which helps in improving your reputation by timely responding to reviews. This can increase the trust and thereby increase in occupancy by responding to reviews.

New CRM strategies are devised to establish and better the customer relations. SENTIMATCH tries to enhance the user experience. Data is aggregated from different sources and specific information is extracted on thousands of properties and added to the search experience. It tries to add

value to multiple independent reviews by automatically analysing the sentiment of each review. By viewing sentiment in past reviews, users can make faster and better decisions. The guided navigation and sentiment analysis capabilities enable customer satisfaction

The data collected over a time period are classified whether it is relevant to be checked or not. This increases the accuracy of the system.

5.6.5 Impact of Social media findings and events on business

The navigation patterns really affect the business. For e.g.

- How many people are just visiting the site?
- How many people are clicking and going to the next level?
- How do they come to the site?
- Did they come by search, typing the URL or arbitrarily?
- What is the exact word used?
- Where did the user go from here?
- Is the user coming back repeatedly?

These features help in predicting whether the person is really interested or not. This also helps to know whether it is activity planning, price factor etc. This is analysed in detail in the later part.

5.7 Data Analysis

Our application starts crawling and collects data. This will scrape reviews from various identified sites and save them as a CSV file. Duplicates and ambiguous data are removed. This data is run through the application. The CSV file created is uploaded to train our classifier. Headers and index columns are removed. Keywords are identified. It is classified by the point of interest, the person of interest and other modifiers. Co-relation is built in tree model based on ontology and taxonomy. Rules are applied. Classification is done and scores are assigned. The classification process helps in identifying mood and hence sentiment analysis. Machine learning algorithm is trained. Model is tested. Training samples are added to classify, retrain and improve results. This gives only basic classification. We need to have machine learning sentiment classifier, to understand different aspects of reviews and to really get insight from the customers.

The reviews are scrapped from trip advisor and booking.com for training data for the entity classifier.

The aspects are divided into positive and negative. The text contents are annotated and divided into small entity units. For fine-tuning, manual intervention (administrator reviews) is required and checked for feedback and accuracy of classification.

5.8 Result Analysis

The opinion mining by sentiment analysis of Facebook post treated here is the guest reviews that will provide insights about the feel of customer satisfaction and dissatisfaction, which is otherwise not explicitly known by just examining the numerical scores of emotions. Very many industries especially in hospitality and healthcare ask the customers to rate in stars, a kind of numerical rating in addition to providing comments. The numerical rating is just a comparison, but it is essential that managers use text analytics to understand the underlying customer sentiments. Such an approach will provide better insights into quality and process improvement.

Three types of online customer review datasets are collected for our system performance.

Data Set and File Descriptions

1. The first set is taken from the available corpus of TripAdvisor dataset.
2. 200 reviews were extracted from Eco-hotel Data Set (UCI Machine Learning Repository [http://archive.ics.uci.edu/], where there were more than 400 reviews. The subjective lexicons and semantic orientation from all the positive and negative sentences were extracted from the experiments done on the data set.
3. 493 reviews were crawled by the scraper for testing our tool developed.

The subjective sentences were processed for semantic orientation by taking the contextual features and using the SentiWordNet for the semantic score. The weight is calculated for the final opinion orientation. The results were evaluated by using precision and recall. The table shows the overall accuracy of our proposed method.

The various files used are :

- train.csv - the training set
- test.csv - the test
- setdestinations.csv - hotel search latent attribute
- sample_submission.csv - a sample submission file in the correct format

All the datasets were processed to remove the noise and clean up the special characters and symbols and check for spelling mistakes; furthermore, we apply the POS tagger and classify the sentences into subjective and objective. The travel review data has already been processed for positive and negative sentences. Only the subjective sentences were taken for further processing to find the semantic orientation at the individual sentence level. Figure 5.2 shows the classification of subjective and objective sentences (taken from our proposed system) for the hotel review datasets. The review data is classified as positive, negative or neutral as per the opinion orientation of comments. The system achieved an accuracy of about 93% for the feedback level and about 86% at the sentence level. So, the rule based system with the lexical system performs better than machine learning

and word level sentiment analysis.

Classification Approach has the following steps

- Data pre-processing
- Text pre-processing
- n-grams TF feature extraction
- unigrams TF-IDF feature extraction
- train and evaluate models
- Finalise the unigrams TF-IDF model

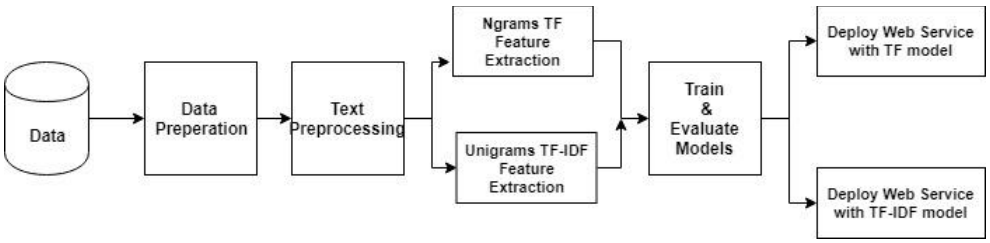


Fig 5.11 Text Analysis

Data Pre-processing

1. A new DATASET is selected from LOCAL FILE

2. Records are labelled with metadata_val module

3. Records with missing text values are removed.

Text Pre-processing

Execute R Script module to specify the required text pre-processing steps.

a) Special characters are replaced with spaces.
b) Duplicate characters are removed
c) Replace numbers with spaces if needed (In the case of some text classification tasks, numbers may be used for training as distinct features.)
d) Convert the text into lower case.
e) Remove stop-words from text using a predefined list.
f) stem the words

Feature engineering - N-grams TF feature extraction

1. Run the Feature Hashing module, specify the Hashing bit size, and specify the n-gram size using the N-grams parameter.
2. Create the Word Dictionary. Extract the set of unigrams (words) to train the text model.
3. The number of documents where each word appears in the text corpus is counted (DF).

TF-IDF Calculation

Higher weight is assigned to words that frequently appear in a corpus. The frequency of occurrence (TF) is used as a feature value. The inverse document frequency (IDF) assigns a lower weight to frequent words. IDF is calculated as the log of the ratio of the number of documents in the training

corpus to the number of documents containing the given word. TF/IDF is important in words that are frequent in the document. This assumption is valid for unigrams, bigrams, trigrams, etc.

Unstructured text data is converted into equal-length numeric feature vectors, where individual feature represents the TF-IDF of a unigram in a text instance. The process is done in the following order.

1. Dictionary is created. The maximum length of a word to be included and the minimum document frequency of a word is calculated.
2. Experiments are run to get the optimal values for the underlying learning algorithm parameters. In the sample experiment, the module will conduct a number of training runs from the parameter ranges.
3. Evaluate the Model module through the results, comparing between the two trained models: N-grams model and unigrams model.
   Based on this evaluation, the best trained model is taken. The corpus used contains unique words (around 10000). The Bag-of-words method is used for feature selection, as well as to reduce the number of unique features. The reduction is done in accordance with the fixed threshold of the frequency of occurrence.

The final feature set consists of 113 entities which are down by 90% decrease. This is inadequate for an accurate prediction.

Training and Testing the Naive Bayes Classifier

Trained on 1500 instances, tested on 500 instances. Accuracy was found to be 0.728.

Most Informative Features

impressive = True	pos : neg = 14.0 : 1.0
outstanding = True	pos : neg = 13.6 : 1.0
offensive = True	neg : pos = 13.8 : 1.0
vulnerable = True	pos : neg = 12.5 : 1.0
Appreciate = True	neg : pos = 14.8 : 1.0
skip = True	pos : neg = 11.7 : 1.0
Pleasure = True	neg : pos = 11.7 : 1.0
astounding = True	pos : neg = 10.3 : 1.0
magnificent = True	pos : neg = 15.3 : 1.0
crazy = True	neg : pos = 9.8 : 1.0

Table 5.3 Most informative features

The 10 most informative features listed above are descriptive adjectives. The two odd words here are vulnerable and skip.

This gave 72% accuracy.

The words above refer to important plot points that signify a good hotel. With simple assumptions, almost 73% accuracy was reached. This is close to human accuracy, which is agreed as 80%.

Precision and Recall for Positive and Negative Reviews

pos precision	0.64269574
pos recall	0.97
pos F-measure	0.7677476
neg precision	0.94877741
neg recall	0.456
neg F-measure	0.63

Table 5.4 Precision and Recall for all the Positive and Negative Reviews

The results can be explained as follows. Word labelled as positive is correctly identified with 97% recall. False negatives are very less or can be neglected in the pos class.

A file classified as pos is only 64% correct, which shows 35% false positives for the pos label. The precision is not good. But on the other hand, a file identified as neg is 96% correct, which means less false positives for the neg class.

Low recall of 45% shows that files which are neg are incorrectly classified.

Low recall of 55% implies false negatives for the neg label. F-measure provides no specifically useful information.

Improving Results with Better Feature Selection

We have used the bag-of-words model for classification. Here every word is independent. Negation cannot be understood by the system. It cannot understand that "not good" is not positive. So training on multiple words is necessary. Neutral words cannot be accommodated into any of this class. So our system should eliminate such meaningless words from feature sets. The modifications, like filter out stop words and include bigram co-locations are added to the feature extraction method. We are using NLTK, which has a list of 128 English stop words in the stop words corpus. Table5.5 shows results for a stop word filtered bag of words :

Accuracy	0.70.86
pos precision	0.650167374005
pos recall	0.974
neg precision	0.958409593496
neg recall	0.469

Table 5.5 Precision and Recall for filtered bag of words

It is seen that accuracy has come down. Accuracy, pos precision and neg precision also went down.

Bigram Collocations

Bigrams are included to improve classification accuracy. The opinions like "not good" is a negative expression, which may be interpreted as positive by the bag of words model. Also in the same way "not bad" which means good may be interpreted as negative as it is two negative words. The challenge here is that significant bigrams are to be identified.

The collocation finder will have two internal frequency distribution for individual word frequencies and another for bigram frequencies. BigramAssocMeasures score individual bigrams, using a scoring function. The collocation correlation of 2 words is measured using the scoring function.

Running iteratively, it was found that using the 100 bigrams from each file produced a good result . This is shown in Teble 5.6

Accuracy	0.82
pos precision	0.749105128205
pos recall	0.93
neg precision	0.91212127
neg recall	0.703

Table 5.6 Bigram Colocation

Table 5.7 shows the Most Informative Features

impressive = True	pos : neg = 15.0 : 1.0
outstanding = True	pos : neg = 13.6 : 1.0
offensive = True	neg : pos = 13.0 : 1.0
appreciate = True	pos : neg = 12.3 : 1.0
('not', 'bad') = True	pos : neg = 12.3 : 1.0
('give', 'us') = True	neg : pos = 12.3 : 1.0
amusing = True	neg : pos = 11.8 : 1.0
skip = True	neg : pos = 11.7 : 1.0
Magnificent = True	pos : neg = 11.7 : 1.0
('not', 'great') = True	neg : pos = 10.6 : 1.0

Table 5.7 Most Informative Features

Here the observation is that when accuracy is up, pos precision, neg recall has increased. It is concluded that significant bigrams increase classifier effectiveness. Without bigrams, precision and recall are less balanced. The differences are data dependent.

Improving Feature Selection will improve your classifier. Reducing dimensionality is one of the single best things you can do to improve classifier performance. Discard the data if it does not add value. Manual

overrun is required for ontology, taxonomy; rule engine, classification and sentiment score for improving performance and accuracy.

It shows the overall performance of the proposed system compared to the machine learning and Hu and Liu methods, taking the feature list as the seed for the opinion orientation. Our system improves the semantic extraction efficiency up to 2% and the opinion sentences extraction up to 10%.

Knowledge base for sentence structure and contextual information

The Knowledge base contains SentiWordNet, WordNet and predefined intensifier dictionaries for domain independent polarity classification for positive, negative and neutral opinions. Sentiment words are usually classified into, positive and negative categories. For this purpose, we extract the semantic score of each opinion word using the SentiWordNet dictionary containing the semantic score of more than 117662 words. Then, we check the structure and associated words (which affect the weight of the opinion word) in the sentence and update the polarity accordingly. The main aspect of this work is a knowledge base for the contextual information of each word in a sentence, which really modifies the strength of the opinion. The knowledge base (calculates semantic strength for each sentence) contains negation words, enhancers, reducers, model nouns, context shifters and other intensifiers with their semantic scores. Negation: Negation words reverse the polarity of opinion words by checking their position in a sentence. The words are (Not, Never, N't, Doesn't, Can't, Nor, Don't, Wouldn't, No, etc). The result will be the opposite if the system fails to

recognize the negation word. So, for the recognition of the semantic expression in the sentence, we use the word sense disambiguation to extract the exact or nearest semantic score of the opinion expression.

5. 9 Converting Visitors to Customers

The website of the eco-resort is visited through direct search, referrals, social media, direct channel and paid channels. Conversion rate from visitors to customers depending on the user experience of each visitor, from these said channels. The company promotes its blog through its Twitter handle. It has a Facebook page promoting its main site.

Clickstream Data is the data trail or impression which a user leaves behind while surfing the Internet. A clickstream is a record of a user's activity on the Internet. This includes the record of the visit of the user to every Web site as well as every page of the Web site. The time stamp helps in getting the information on duration of the visit of the user on a page or a site. The order of the pages visited and interactions can also be tracked. This track can be done by ISP or individual Web sites.

Captured on weblogs, clickstream data information includes the browser name, type of device used (laptop, desktop, tab, mobile devices), Day, Timestamp, IP address, URL, etc. This data helps in understanding the individual users browsing pattern and can be used to analyse and infer patterns, trends, and pathways to a given page or a product. Machine understands and learns the entry point from where the visitor reaches one

page, where he is exiting and everything in between.

5.10 Extracting information from Clickstream data

Clickstream data of a visitor may be one hit or several hits during a visit. When you take a time period, we have a collection of visits. The data at the visitor level need to be organised.

Site visitor's dataset can be structured in rows and columns with each row consisting of a unique subject and each column corresponding to information about that subject. For doing customer-based analysis, customer based data set is considered. A chart with the following details is created for analysis.

2014-07-26,12:54:16,34.09.130.15,98.48.225.22,GET,80,200,10801,"http://www.banasura.com/"

2014-07-26,20:55:15,98.47.227.21,98.46.220.42,GET,80,200,10801,"http://www.banasura.com/luxury-resorts-in-wayanad-kerala"

2014-07-26,22:18:01,12.41.114.23,98.45.225.98,GET,80,200,10801,"http://www.banasura.com/kerala-cuisine-dining-resort-in-wayanad"

2014-07-26,13:15:06,98.46.230.79,98.47.126.99,GET,80,200,10801,"http://www.banasura.com/honeymoon-holiday-package-in-wayanad-kerala

2014-07-26

22:15:06,98.45.226.19,98.47.214.50,GET,80,200,10801,"http://www.banasura.com/wayanad-nature-resorts-package-tariff"

2014-07-28,22:16:22,11.11.111.11,98.47.214.50,GET,80,200,10801,"http://www.banasura.com/trekking-packages-in-wayanad-kerala"

The data may be organised as

Visitor id Total visit Page Views Time Spent

Different statistical models and data science techniques can be applied to a customer dataset. Clickstream data analysis is a powerful and cost-effective tool benefiting businesses in the following ways -

- Understand the profile and characteristics of different visitors, which pages web visitors are visiting and in what order.(Whether they are looking at price and going to the booking page?)
- This helps in knowing how to engage with them (How often are they coming?)
- Understand the level of brand awareness of individual customers (whether they are coming directly to the page or through some other referrals)
- Target customers with individualized, relevant offers (like weekday offers, group offers)

- Anticipate which customers are most likely to convert and statistically

- Understand the content that visitors are most involved with and how content engagement drives high-value visits (Day specialities, walking tour, plantation tour, etc.)

The user experience for different channels helps the site to maximize conversion rates for organic search, social, and direct traffic.

A typical scenario can be explained as follows:

- A visitor from any part of the world wants to visit Kerala, specifically Wayanad and enjoy the natural beauty. Thanks to some excellent feedbacks posted in different locations, the name has become familiar and so instead of searching, the visitor types in banasura.com directly into the browser.

- For analysing the clickstream data, user's activity on a web page or application is captured. Clickstream R package is used to analyse the Clickstream data and the package uses Markov Chain modelling.

- A “session” is defined for a web page as the time from entry of the visitor till he is logout or timeout. Similarly, the time between two subsequentt application start events can be considered as "session" for an application.

- The collected event log is transformed into clickstream data by

identifying the events performed by the same user and clustering them. This is further subdivided into groups of events, based on the performance during the same session.

The dataset is transformed as session followed by events performed by the user in that session

Ses1 E8

Ses2 E14 E4 E8 E11 E12

Ses3 E14 E4 E8 E11 E12

Ses4 E14 E4 E9 E8 E9 E8 E11 E12

Ses5 E14 E4 E9 E8 E11 E24 E9 E9 E8 E1 E14 E4 E8 E11 E12

To model the above problem of clickstream analysis, we utilize Markov Chains, which work with sequential data. The Markov process is a stochastic process that satisfies the Markov property of memorylessness. A Markov chain is a Markov process in either discrete or continuous time, with a countable state space.

Here the process Xn takes the state: Mn from a finite set: M at each time:n. The order of a Markov Chain is derived from the number of recent states, on which the current state depends. So the zero-order chains imply that the probability of being in a state in the next step is independent of all previous states. The higher-order Markov Chain introduced by the Raftery [110] lead

to more realistic models. With the higher order Markov chain, the parameters for the representation increase exponentially.

5.10.1 Fitting a Markov Chain

As mentioned before, at this point, our dataset looks like:

Ses1 E8
Ses2 E14 E4 E8 E11 E12
Ses3 E14 E4 E8 E11 E12
Ses4 E14 E4 E9 E8 E9 E8 E11 E12
Ses5 E14 E4 E9 E8 E11 E24 E9 E9 E8 E1 E14 E4 E8 E11 E12

We choose third-order Markov Chain on the above data. The reasons for choosing are:

- The manageable number of parameters needed for the chain's representation (optimum computing power).
- Half of the clickstreams consist of as many clicks as the order of the Markov Chain that should be fitted. (a thumb rule)

Fitting the Markov Chain model, gives the transition probabilities matrices and the lambda parameters of the chain, for each one of the three lags, along with the start and end probabilities.

Start and end probabilities correspond to the probability that a clickstream will start or end with this specific event. The transition probability matrix can be represented as a heat map, with the y-axis representing the current state and x-axis representing the next one.

5.10.2 Click Prediction

The next click or the final click (state) of a user can be predicted in a clickstream, by the pattern they have been following. This data-driven personas or profiles can be constructed which incorporate users' behaviour. The probability of transition is calculated along with the click prediction.

Consider the event state E14, the next probable transitions are:

E14----(0.4957)--->E4----(0.5)----> E11

This helps in finding the common use case scenarios of an app or a web page.

5.10.3 Clustering Clickstream Data

Due to the complexity of websites or applications, same clickstreams are difficult to occur. A user can follow many different paths. This results in a large number of different clickstreams difficult to analyse. The similar clickstreams and user profiles are grouped. This helps in identifying customer segments and communities with similar interests. We perform k-means clustering with two centres. The clustering of clickstreams is based on the actions of the user during a session. The cluster interpretation mostly requires a deep understanding of the data and domain expertise.

Because of the above said challenges, representing as sequential patterns are considered instead of transition probabilities. This permits getting even the patterns, which occur a small number of times in the user's clickstream data.

The sequence of interactions while the user is searching or browsing is analysed. Frequent Sequence Mining is used to observe and detect the digital footprints shared between the events in a specific order. For frequent sequence mining, the SPADE (Sequential PAttern Discovery using Equivalence classes) algorithm [102] is used. R package arules is used for creating and analysing item sets and rules. Add-on for arules, arulesSequences package is used to handle and mine frequent sequences. SPADE algorithm mines frequent patterns. For mining, the algorithm starts by computing the frequencies of sequences with one item, then in the second step with two items and so on. SPADE algorithm extracts all pattern sequences with minimum support defined. For a given sequence pattern S, we can predict the next click by searching for the pattern sequence with the highest support that starts with S.

Two models using Markov Chains and the SPADE algorithm for mining clickstream sequence data are explored. The model helps in

- Creating personas for the footprints of the customers on a web page.
- Predict the next actions based on previous actions.
- Extract frequent sequential patterns.

The business can understand, explore or predict the visitors' interest through the navigation patterns of a website or application. But the challenge here is that a lot of human intervention is required.

5. 11 Findings

Customer reviews about the eco-resort shared at different social media channels are mostly unstructured comments. Monitoring and mining customer conversations leads to clues which are triggers for customer behaviour. NLP methodologies make sense out of ambiguities in human language. The tool SENTIMATCH analyses and provides actionable managerial insights to improve operations. SENTIMATCH gives the most insightful content to analyse. This can be considered as a reputation management program, which helps in improving your reputation by timely responding to reviews. This can increase the trust and thereby increase in occupancy by responding to reviews.

The results of our methodology is encouraging, with scope for further improvements. The fifth research question number is addressed here. To achieve our research objective and to illustrate the methodologies, the experiment on the public data set reveals the following

- The reviews posted by the customers help the business to know what their guests feel about the property. Overall and individual sentiment of each review helps in making operational and strategic changes by the management
- Information gives insight into how the resort can improve their operations and meet customer expectations. The feature selection helps to refine the aspects which customer is addressing.

- The actionable items to be integrated into CRM for driving more traffic to the site.

This is the reputation management program which helps in brand positioning by detecting the sentiments of the customer. Immediate corrective action is taken if any negative feedback is detected. This answers the sixth research question.

The casual visitors are classified by the rule engine as interested or not interested and that lead is passed on to CRM, for mailing or another contact for conversion. The conversion rate is found to be between 50% to 60%. It is also found that 20% of customers are through WOM, 30% through aggregators like TripAdvisor, Agoda etc. There are many overlaps also.

The evaluation placed the accuracy of classification based on Sentiment Analysis to be 72%.

5. 12 Further Enhancements

In the current setup, the data is from the Facebook post and from sites like TripAdvisor, Agoda, only. This can be extended to other social media posts including tweets, blog posts, forums, aggregators, etc.

The conclusion here is that there is a need for constant monitoring of the customer sentiments, get a better understanding of what customer is looking for in the dynamic updation of the contents. These changes should engage visitors and keep their attention.

In this experiment, the domain explored is a subset of tourism and the data is collected. A wealth of unstructured opinion data on various domains are available across the social media, which are yet to be tapped. The crowdsourcing and opinion mining is set to change the way the business is conducted. Analytics provides the data of this transaction for future predictions. This helps marketers in making decisions to improve the customer experience. The bottom line is increased sales.

…….&&.....